MODEL-BASED ENHANCEMENT OF MOVING FACIAL IMAGES

XIAOMENG PING

Model-Based Enhancement of Moving Facial Images

By

© Xiaomeng Ping, B.Eng.

A thesis submitted to the School of Graduate Studies
in partial fulfilment of the requirements for
the degree of Master of Engineering

Faculty of Engineering and Applied Science
Memorial University of Newfoundland
August 1998

St. John's Newfoundland Canada

Abstract

This thesis investigates the application of 2-D object modeling to image enhancement and restoration. In particular, the topic of the thesis is focused on the recovery of accurate facial images from noisy and blurred videos. An immediate application of this is the improvement of security videos for identifying people, but there are other contexts, such as film restoration, where views typically fixate on faces, and there is therefore reason to give special importance to improving that part of the image.

When several video frames of a static scene are available, noise reduction can be accomplished simply by averaging all the frames. In practice, movement of the objects (or the camera) often prevents this. For example, a security video will probably provide a frame sequence of an individual in motion. Our strategy for enhancing facial images is to employ *warping* to eliminate these variations and to produce a uniform image sequence which will be involved in *frame averaging*.

A 2-D face model which consists of a set of landmarks that describe the main facial feature points is used in the thesis. The face model is matched frame-by-frame to the sequence by a human operator. Given the landmarks, warping is performed to transform all the noisy and blurred frames to the same target frame. Then temporal averaging among the transformed resulting frames is applied to produce an enhanced result. Finally, spatial filtering for blur removal is optionally applied.

To test the effectiveness of this strategy, some experiments are designed and performed. The results demonstrate that the supervised object-based temporal filtering strategy is simple and effective for enhancement of faces in videos.

Acknowledgements

There are many people who have contributed to the final form of this thesis.

I own an enormous debt of gratitude to my thesis supervisor, Dr. John A. Robinson, for being a constant source of ideas and inspiration, which led me to many interesting and fruitful lines of investigation in these two years.

I own special thanks to Li-Te Cheng, especially for his great work on MCLGallery, which provides a so convenient programming environment for my user's interface. I would also like to thank Qing Song, who is working on human faces too. I enjoyed sharing tons of paper and other resources with her.

I own a big thanks to everyone at the MCL lab who volunteered to be my face candidate and participated in my subjective tests. They are Charles Robertson, Yan Shu, Rahim Pira, Moorthy Manoranjan, Xincheng, Qing Song, and Li-Te Cheng.

Finally, I am grateful to my parents for their overseas constant moral support.

The research reported in this thesis was supported by Graduate Studies of Memorial University of Newfoundland, National Sciences and Engineering Research Council of Canada (NSERC), Newtel Communications (Newtel) and Northern Telecom (Nortel).

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Problem Statement and Applications

In our daily life, video is used for a variety of evaluative and interpretive purposes from surveillance cameras used to monitor places such as banks and shopping establishments, to sports events where a referee may use video footage to assist in making difficult decisions. This brings about the importance of the image sequences obtained from video films as a scientific tool.

However, image recording systems are not perfect. As a result, some external components that interfere with the visual perception of the real world contaminate all images to various levels. These undesired components are called *noise*. Sources of noise include electronic noise, photon noise, film-grain noise, and quantization noise. In addition, images may be blurred due to camera misfocus or relative object-camera motion.

In this thesis, the means to recover accurate *facial* images efficiently from noisy and blurred videos are explored. An immediate application of this is the improvement of security videos for identifying people. The visually enhanced result can also be used in automatic face identification systems which rely on the input image quality. This smooths away the difficulties for security teams and law enforcement officers when only poor quality video is available. In addition, there are other contexts, such as film restoration, where viewers typically fixate on faces. The ability to enhance the facial image stream is therefore tremendously useful in many such situations.

1.2 Mathematical Model

In our work, linear degradation due to noise and image blurring is studied. This is done because of the relatively simple linear degradation model which can be used,

$$\mathbf{Y}_i = \mathbf{H}\mathbf{X}_i + \boldsymbol{n}_i, \qquad i = 1,2,3,\ldots, M$$

where $\mathbf{X}_i$ is the original undegraded face data for the ith frame in a sequence, $\mathbf{H}$ a linear degradation operator, which is space and frame invariant, $\mathbf{n}_i$ an additive noise term, $\mathbf{Y}_i$ the corresponding observed face data and M the number of frames we have obtained. Therefore, the facial image enhancement problem is then simply stated as to estimate $\mathbf{X}_i$ from $\mathbf{Y}_1$, $\mathbf{Y}_2$,$\cdots$ to $\mathbf{Y}_M$ with some knowledge or just assumptions about $\mathbf{H}$ and $\mathbf{n}_i$.

1.3 Approach Description

Existing software for security purposes provides general image processing techniques, such as image editing, adjusting brightness and contrast, zooming, smoothing and sharpening. There is no effective method to handle the degradations caused by extreme noise and blur. An efficient approach to handle this situation is developed by working directly with object information (a face in our case) and the associated temporal context, i.e., the past and future frames.

This thesis' strategy for enhancing facial images is to fit face models, as used in animation [1][5], videophone compression coding [18] and face recognition [19], to the image of the face in several frames. Model point correspondences are then used to filter the images temporally. For example, the tip of the nose is identified in a series of images where it is visible but perhaps oriented in different directions. All the different instances

of that point are considered as projections of a single model; they can be averaged (perhaps with weighting) and then displayed as the enhanced nose tip.

Specifically, a two dimensional face model rather than a three dimensional model and a human supervised model fitting procedure is used, i.e., the 2-D model is overlaid, frame-by-frame on the video, by a human being. The strategy is based on the following considerations:

First, in the main applications of security and film restoration, real time processing is unnecessary. Hence, it is possible for us to perform this enhancement offline.

Second, the human-supervised procedure in successive frames is the only approach to handle the images with very poor quality where existing face detecting and locating algorithms are not able to work.

Third, despite the drawback of misrepresenting depth while face turns, a 2-D face model has the obvious advantage in a model fitting procedure compared with a 3-D face model because no depth information needs to be estimated.

Therefore, when a fully automatic system has difficulty dealing with the images in extreme noise and blur, this strategy is the most and perhaps the only practical solution to such a problem.

1.4 Structure of Thesis

This thesis describes a manual 2-D face model fitting procedure applied to a sequence of video images, then temporal averaging between the images to produce an improved result. Chapter 2 gives some background knowledge in this area. Chapter 3 describes the face model, its landmarks, the mechanism for warping the model onto an

image, and the technique to balance the shadings on a face. Chapter 4 shows the method applied to image sequences with varying degrees of noise and blur degradation, and resulting enhanced images, as well as the filtering applied for blur removal after frame averaging. Chapter 5 shows results on simulated security videos. Chapter 6 investigates whether the method is sensitive to operator (user) variation. Chapter 7 displays the recovery results for the entire real videos. Chapter 8 concludes the thesis and briefly describes future work.

Chapter 2 Background

In this chapter, some existing standard techniques for image enhancement which are part of the overall system proposed in this thesis (e.g. histogram equalization and high-pass filtering) or compared with the proposed system (e.g. the low-pass filtering) will be reviewed.

2.1 Methods for Image Enhancement

The goal for image enhancement is to accentuate or sharpen image features (e.g. edges, boundaries and contrast) and to attenuate noise for subsequent analysis or for display. This process does not increase the inherent information content in the image. It rearranges the information in a way suitable for human vision so that features can be detected easily.

In this section, traditional enhancement methods including contrast enhancement, spatial filtering and temporal filtering developed from spatial domain techniques will be discussed. The spatial domain refers to the aggregate of pixels composing an image, and approaches in this category are based on direct manipulation of the pixels in one image or some correlated images.

A general expression for the spatial methods can be described as

$$v(x, y) = f[u(x, y)] \tag{2.1}$$

where $u(x, y)$ is the original image, $v(x, y)$ is the processed image, and f is a function on u.

The spatial techniques can be broadly classified into three operations according to the range of the definition of f. The first one alters the contrast range occupied by pixels

in an image by using some specified algorithm to determine a new value for each pixel. This is called *a point operation*, where f is defined for each particular pixel positioned at (x, y). Section 2.1.1 is concerned with this operation. The second operation involves context-dependent operations that alter the reflectance value of a pixel according to its relationship with the grey levels of the other pixels in the immediate vicinity. This is *a neighborhood operation*. Filters given in section 2.1.2 belong to this operation. The third operation is to perform f on some successive frames. This is called *a temporal operation*, Filters in 2.1.3 are examples of this operation.

2.1.1 Point Operation and Contrast Enhancement

The contrast of an image is subject to the distribution of pixel grey values. If the grey values are concentrated near a certain level, the image has low contrast. This occurs often due to poor or nonuniform lighting conditions or due to nonlinearity or small dynamic range of the imaging sensor. On the other hand, a wide range of grey levels gives a high contrast image.

Histogram

For a digital image with grey levels in the range $[0, L\text{-}1]$, the occurrence of the kth grey level is easily obtained by the following calculation:

$$p(k) = \frac{n_k}{n}$$

where n_k is the sum of pixels in the image with grey level k, n is the total number of pixels in the image. A plot of this function versus k (where $k = 0,1,2, \ldots, L\text{-}1$) is called a *histogram.*

Histograms provide useful information on both the global contrast distribution of an image and the possibility for contrast enhancement. Generally, contrast manipulations within the image will be reflected in corresponding changes in the histogram.

Histogram Equalization

Histogram equalization is a point operation where a given grey level $u \in [0, L-1]$ is mapped onto another grey level $v \in [0, L-1]$ according to a transformation $v = f(u)$, which is independent of the pixel position. Normally, the function f should be increasing to retain the brightness order of the original image, that is, $f(u_1) \le f(u_2)$ if $u_1 < u_2$.

The mapping function for histogram equalization is based on the histogram and is defined as:

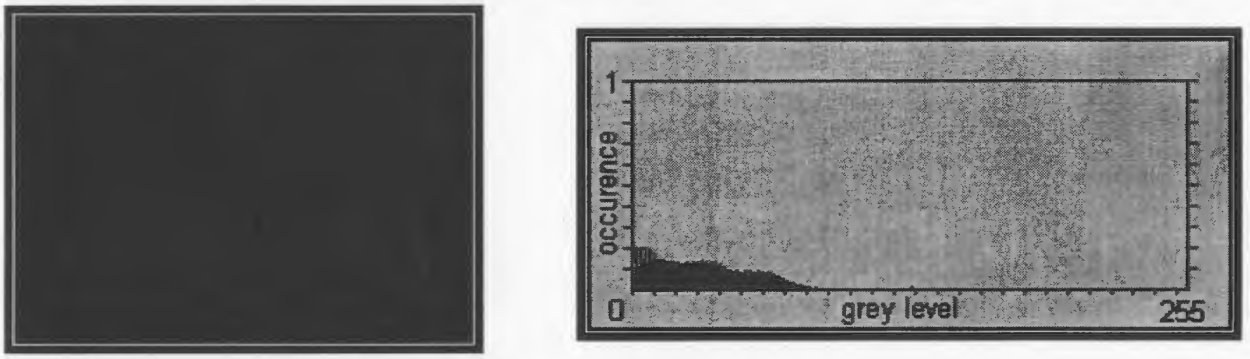
$$v = f(u) = L \times \sum_{j=0}^{u} \frac{n_j}{n} = L \times \sum_{j=0}^{u} P_j \qquad 0 \le u < L$$

where L is the grey level range (for example, $L = 255$ for a 256 color image), n is the total pixel number in the image, n_j is the sum of pixels with grey level j, and P_j gives the occurrence probability for the grey level j.

It can be easily verified that the mapping function is increasing and approximately an equal number of pixels for each grey level implies a more uniform histogram is expected after histogram equalization. The increased brightness gradient spreads out the dynamic range of grey levels, and, consequently, brings about an improvement of image contrast.

Here "approximately" is used because a perfectly equalized histogram can not be obtained due to discontinuities of the equalizing function *f*. As an example, Fig.2.1 shows the frames and their corresponding histograms. Before equalization, the histogram has a relatively narrow shape, which indicates little dynamic range and thus corresponds to an image having relatively low contrast. After equalization, the histogram has a significant spread, corresponding to an image with a relatively high contrast.

(a) Before Histogram Equalization



(b) After Histogram Equalization

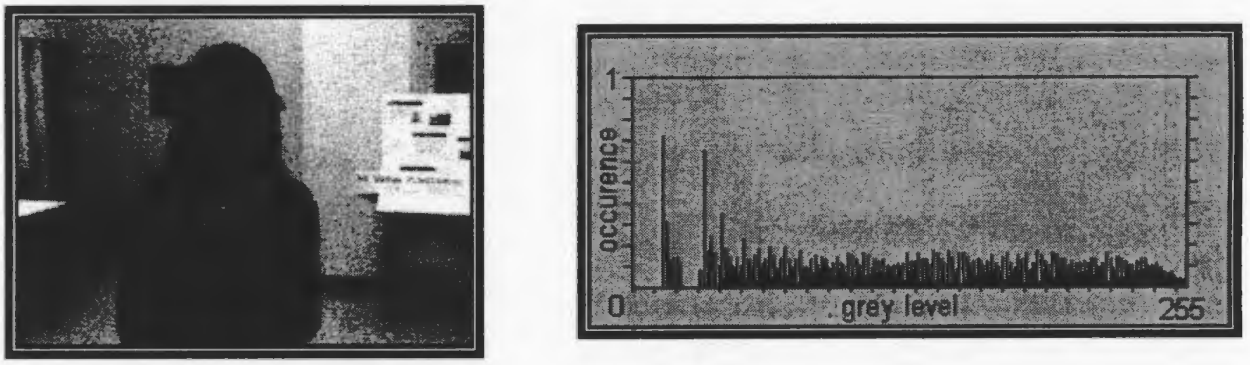


Fig.2.1 Frames and Histograms

2.1.2 Spatial Filtering

Filtering is a process to generate a correcting effect on the image by removing certain components of the undesired signal while retaining others at the same time [26][27][28].

In spatial filtering, the new grey level for each pixel on the input image is determined by its neighborhood pixels.

Low-Pass Filter

Low-pass filtering refers in particular to situations where the high frequency content of the image is attenuated and low frequency content is approximately constant.

Normally, image information is represented in a regular and uniform fashion without large intensity variations between neighbouring pixels, which map to low frequency components of the image, while noise has a broadband content. Therefore, a low-pass filter can be used for reducing noise.

Besides noise, sharp edges and details in an image also corresponds to high frequency information, which will be weakened by low-pass filtering. This leads to another application for low-pass filters: smoothing images, which is also the inevitable drawback for the application in noise removal.

Next, three kinds of low-pass filters involved in the work of this thesis, weighted-average filter, median filter, and MMSE filter, will be described.

1. Weighted-Average Filter

A weighted-average filter is an operation to replace the grey level of each pixel in the input image by a weighted summation of its neighbors within a specified mask.

In this thesis, a weighted-average filter is utilized to create blur degradation on clean images. During the experiment discussed in Chapter 4, blurring effects in three varying degrees are desired: *slight blur*, *moderate blur* and *heavy blur*. This is satisfied by adjusting the spatial masks of the weighted-average filter. The weights and sizes of masks for these filters are defined in Fig.2.2. Please note that heavy blurring is obtained by operating moderate blurring three times successively.

$$\frac{1}{6}\times\begin{bmatrix}0&1&0\\1&2&1\\0&1&0\end{bmatrix}$$

(a) 3×3 mask (for slight blurring)

$$\frac{1}{100}\times\begin{bmatrix}1&2&4&2&1\\2&4&8&4&2\\4&8&16&8&4\\2&4&8&4&2\\1&2&4&2&1\end{bmatrix}$$

(b) 5×5 mask (for moderate blurring)

$$\frac{1}{100}\times\begin{bmatrix}1&2&4&2&1\\2&4&8&4&2\\4&8&16&8&4\\2&4&8&4&2\\1&2&4&2&1\end{bmatrix}+\frac{1}{100}\times\begin{bmatrix}1&2&4&2&1\\2&4&8&4&2\\4&8&16&8&4\\2&4&8&4&2\\1&2&4&2&1\end{bmatrix}+\frac{1}{100}\times\begin{bmatrix}1&2&4&2&1\\2&4&8&4&2\\4&8&16&8&4\\2&4&8&4&2\\1&2&4&2&1\end{bmatrix}$$

(c) 5×5 mask (for heavy blurring)

Fig.2.2 Masks for Weighted-Average Filters

2. Median Filter

Median filtering is a nonlinear procedure that is usually used for noise reduction. The theory is that noise is always represented in an isolated fashion while image information corresponds to blocks with relatively large size.

For a two dimensional median filter, the grey level of each pixel in an image is replaced by the median of the grey levels in a mask centered on that pixel, i.e.,

$$v(x,y) = median\{\ u(x-m, y-n),\ \ (m,n) \in W\}$$

where $u(x,y)$ is the input, $v(x,y)$ is the output and W is a suitable selected mask.

The median filter is effective particularly for images degraded by binary noise, i.e. the noise can be removed while the image details are approximately preserved. But it performs poorly for Gaussian noise and heavy noise when the number of pixels effected by noise in the mask is greater than half of the number of pixels in the mask.

In this thesis, recovered results by the median-filter are used for comparison with the new approach proposed in this thesis. Please refer to Section 5.6 for details.

3. MMSE Filter

The aim of minimum mean-square error (MMSE) filter is to remove noise while keeping details in an image [28]. Different from the filters discussed previously, the MMSE filter has adaptability in the sense that its properties are modified by the local image statistics.

If $u[x,y]$ and $v[x,y]$ represent the grey level intensity for the considered pixel at $[x,y]$ in the original image and the filtered image respectively, the general form of MMSE filter is defined by the following equation:

$$v[x,y] = u[x,y] - \frac{\sigma_n^2}{\sigma_l^2[x,y]}(u[x,y] - m_l[x,y])$$

where σ_n^2 is the noise variance of the original image, $\sigma_l^2[x,y]$ and $m_l[x,y]$ are the local grey level variance and local grey level mean for the pixels within the mask centered at $[x,y]$ respectively.

As indicated from the equation, the output of the MMSE filter consists of three parts: the original grey level $u[x,y]$, some original value to be subtracted, and some local mean to be added. The specific weights for the subtraction and addition are adjusted by the noise to local variance ratio. For the regions varying slowly in the uncorrupted image, the ratio will increase to 1, which results in the local average being the main part of the filter output. For the regions with sharp edges, the ratio will decrease and the pixel will pass the filter almost without change.

Examples by MMSE filter are given in Section 5.6 for comparison with other approaches.

High-Pass Filter

On the contrary, high-pass filtering will attenuate the low frequency content of the image and keep the high frequency content constant. It is used for contour enhancement because high frequencies describe edge information.

The typical 3×3 mask of a spatial high-pass filtering is usually in the format shown in Fig.2.3 (a), where m and n are any positive integers. The mask in Fig.2.3 (b) (where m = n = 1) which leads to the best results is employed in the experiments in this thesis when an image sharpening procedure is desired.

$m \times$

−1	−1	−1
−1	8	−1
−1	−1	−1

$+ n \times$

0	0	0
0	1	0
0	0	0

(a) m and n are any positive intergers

−1	−1	−1
−1	9	−1
−1	−1	−1

(b) m = 1, n = 1

Fig.2.3 High-pass Filtering Mask

From the previous discussions, undesired blurring occurs along with noise reduction in a low-pass filtered image. Similarly, in a high-pass filtered image, undesired noise will be emphasized as well as contours. It may be possible to reduce the noise with more sophisticated filters such as laplacian-of-a-gaussian filters [26]. As for the spatial filters, these are inevitable deficiencies of the low-pass filter used for noise removal and high-pass filter used for image deblurring. Therefore, the choice of spatial filter to enhance the both noisy and blurred images is a compromise between the desired quality and the accompanying disadvantages.

2.1.3 Temporal filtering

All the filters described in 2.1.2 are operations on pixels within a single frame, where only spatial information is used. As just mentioned, the fatal drawback for low-

pass filters is that there is a tradeoff between noise reduction and spatial blurring of the image detail.

Please examine the general expression for spatial filtering from equation (2.1) $v(x,y) = f[u(x,y)]$ in page 5 again. For all the filters discussed in 2.1.2, the operator f is defined over some neighborhood of (x, y). A temporal filter will be obtained if f is also performed on a set of correlated images. Therefore, temporal filters are 3-D filters with two spatial coordinates and one temporal coordinate, i.e., not only the spatial correlations, but also the temporal correlations between the frames are utilized in such filters.

In video processing, temporal filtering may provide more advantages than spatial filtering. In principle, temporal filters can avoid spatial and temporal blurring by exploiting the essential distinction between noise and image information in the image sequence, i.e., the noise is uncorrelated among frames, while the image is highly correlated from frame to frame.

In this section, schemes based on temporal filtering for both still images and moving images will be reviewed.

Frame Averaging for Still Pictures

The simplest form of temporal filtering is *frame averaging*, where pixels occupying the same spatial coordinates in consecutive frames are averaged. Frame averaging corresponds to maximum likelihood estimation under the assumption of additive white Gaussian noise (AWGN). It can be verified that it reduces the variance of the noise by a factor of the number of samples.

Assume the grey level intensity of a particular pixel with co-ordinate (m,n) in the ith image of a noisy sequence is:

$$u_i'(m,n) = u(m,n) + \eta(m,n)$$

where $u(m,n)$ is the actual grey level and $\eta(m,n)$ is the AWGN with zero mean and σ_η^2 variance. The average grey level at (m,n) by M frames is simply:

$$v(m,n) = \frac{1}{M}\sum_{i=1}^{M} u_i'(m,n)$$

This yields the representation of the statistical expectation and variance of $v(m,n)$ as

$$E\{v(m,n)\} = u(m,n)$$

$$D\{v(m,n)\} = \sigma_{v(m,n)}^2 = \frac{1}{M}\sigma_{\eta(m,n)}^2$$

Obviously, the variance of $v(m,n)$ will decrease and $v(m,n)$ will approximate the original image $u(m,n)$ as the frame number M increases. Fig.2.4 shows the corresponding probability density function of the original frame $u'(m,n)$ and the averaging result $v(m,n)$.

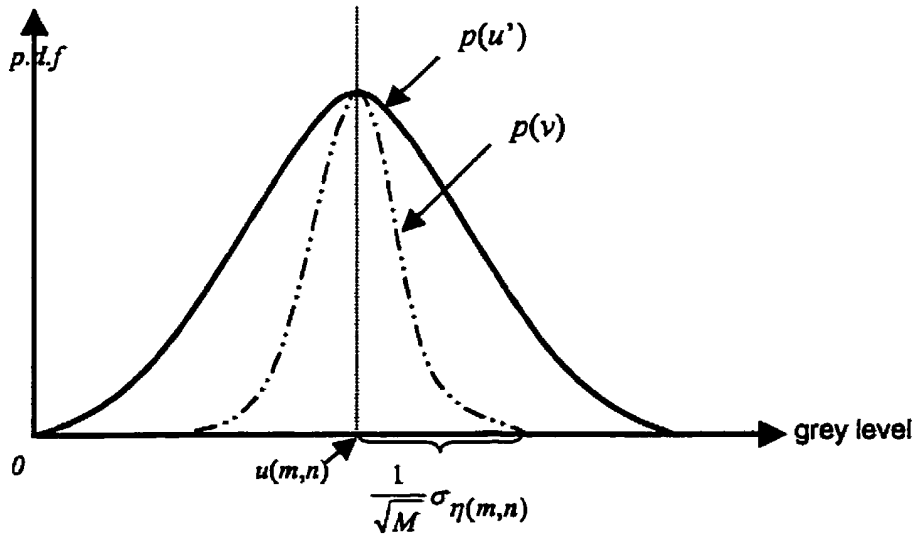


Fig.2.4 Probability Density Function of $u'(m, n)$ and $v(m, n)$

Given enough frames, frame averaging has the distinctive advantage of reducing the noise to practically negligible without loss of spatial image resolution. However, frame averaging is suitable only for when the same object is static throughout the video sequence. In practice, the frame sequence usually does not satisfy this requirement because of the movement of the object. For example, a security video will probably provide a frame sequence of an individual whose movement produces different sizes, orientations and impressions.

If we apply frame averaging directly on the moving facial images, it may lead to smearing in the moving areas, which is similar in appearance to the lag from a camera. As described later, *warping* can be employed to eliminate these unfavorable factors and produce a uniform image sequence. Specifically, in this thesis, warping is done using a 2-D face model, matched frame-by-frame to the input by a human operator.

Adaptive Temporal Filtering for Moving Pictures

Other approaches to avoid degradation on moving pictures by frame averaging are described in the literature [2][3][4], where the picture is segmented into stationary areas and moving areas by motion detection and the temporal filtering is only applied to the stationary areas. The disadvantage for these systems is that the noise in the moving areas can only be reduced at the expense of image detail.

In [5], improvement is made by exploring the interframe motion information. Assuming the motion trajectory of the object can be exactly estimated, it is reasonable to make the deduction that the high frequency component of the variation for a object point is mainly from noise, because the variations in the image point are mainly due to change in the luminance or orientation of the object, which corresponds to the low frequency

component. Hence by performing a low pass filtering operation on image intensity at a given object point, the noise component can be significantly attenuated, with a minimal effect on the image component.

A straightforward block diagram for such filters takes the form shown in Fig.2.5. It is a first-order recursive temporal filter with motion compensation similar to an interframe differential PCM coding loop by replacing the normal quantizer with a multiplier. In this diagram, u is the input frame and v is the corresponding output frame from the filter. Based on the previous frame intensity of the current picture pixel, the prediction can be made. The prediction error e at the output of the subtractor is mutiplied by λ (assume λ is a constant) less than unity and then added to the predictor to calculate both the pixel value of the output picture and the input of the predictor. In this system, the noise reduction is about $10\log\frac{2-\lambda}{\lambda}dB$ for the stationary areas.

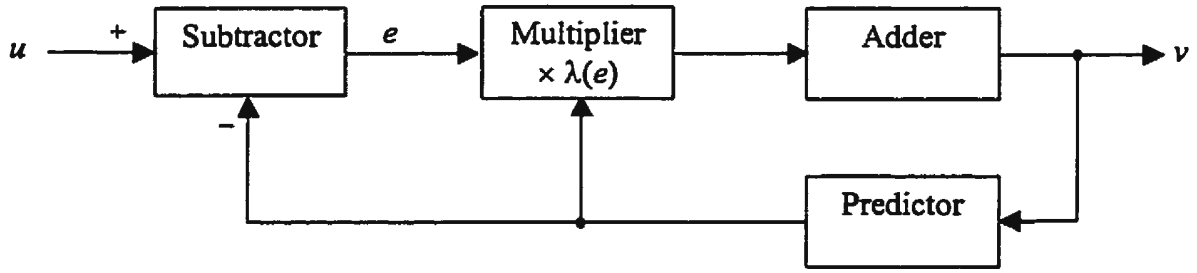


Fig.2.5 First-order Recursive Temporal Filter with Motion Compensation

However, the system has the drawback of introducing artifacts on the newly exposed regions, where accurate prediction can't be obtained and large prediction errors are produced. One modification proposed in [6] is to minimize the degradation effect by

adapting the multiplier coefficient λ to the amount of motion at each pixel based on the assumption that this can be distinguished from noise. Specifically, λ is dependent on the prediction error e such that it is low for small e and close to unity for large e.

From the theory point of view, this filter introduces no spatial degradation for little motion, and it will attenuate large-area spatial interference, such as streaking, provided it does not contain zero-frequency or frame-frequency components.

2.2 Face Models

A face model is a mathematical abstraction of the appearance of a face to some degree of accuracy in a way that makes the model useful for specific applications. Typically, the models are designed to produce meaningful facial images.

Although most faces are similarly structured with the same facial features arranged in roughly the same spatial configuration, significant geometrical or texture differences exist due to individuals, ages, expressions, view positions and common structural features such as glasses, hairstyle or a moustache. Furthermore, unpredictable imaging conditions in an unconstrained environment will lead to even more variability in face patterns because of shadings, noise and blurring effects.

Clearly, one of the most challenging issues in face models is to devise a reliable scheme that can accurately account for the wide range of permissible variations in face patterns.

2.2.1 Applications for Face Models

Typical models of the human face are relevant to a variety of applications, such as computer animation, communications, medicine, face recognition and interpretation. The amount of face detail that the model captures highly depends on the sort of application being developed.

Computer Animation

The first attempts in computer-based facial animation involved key-framing, where two or more complete facial expressions are captured and in between frames computed by interpolation [11]. The immense variety of facial expressions makes this approach extremely data intensive. In 1974, Fred Parke first utilized 3-D computer graphics to simulate human faces in his Ph.D thesis, where a parametric face model [7] was proposed.

A computer-generated face, definitely more familiar for users than the texts, will not only provide a comfortable and friendly working environment, but also has distinct advantage over images of real people in situations when precise and repeatable facial actions are desired. These faces suggest some unique and novel situations for presenting information, especially in a noisy environment. Examples of this type of interaction can be found in walk-by kiosks, ATM tellers and office environments. In the near future, the man-machine interfaces that mimic the way people interact face-to-face are expected. Hence, these will take away the embarrassment some users may feel when facing a boring keyboard.

In this context, the synthesis of facial expressions is important because of the requirement of the harmony of empathy and human emotion with computer generated characters.

Computer-generated faces also have potential application in narration because visual information obtained by speech reading and interpretation of body gestures improves perception of speech [17]. In this context, a face model which incorporates speech synthesis capabilities could prove to be useful for hearing impaired people.

Medicine

The main uses for face models in medicine will be in the surgical and psychological areas. In [10], the effects of corrective surgery on patients can be predicted with the aid of parameterised facial models without undertaking costly and potentially dangerous exploratory surgery. Such applications demand very precise models of particular individuals based on the bone and soft tissue of the head.

Face Recognition and Identification

The face is one of the most convenient and reliable personal IDs for many customer services where access is only for authorized people. That is, the verification of individuals by faces can enhance privacy and confidentiality while reducing fraud. Recognition and identification of faces is also an important aspect in criminal investigations. The system developed by Vision Control Australia and the Victoria Police uses a lot of the knowledge from research on facial modelling along with the experience that police have had with the Identikit and Photofit identification systems [9]. Here, representing the appearance of a wide variety of faces is particularly important.

Communications

Face models are an effective solution to the transmission of scenes which are generally restricted to head-and-shoulder types in teleconferencing or videophones over low-bandwidth channels. In this case, a photo realistic model of the speaker is captured and transmitted to the receiving station where it is reconstructed at low bit-rates to produce a realistic animated image of the speaker's face. More details on the strategy of model-based communication coding will be discussed in the next section.

2.2.2 Face Model Representations

2-D Face Models and Feature Vector

If we look at the face as a 2-D structure, the most simple but also least compact approach is to model a face by the original 2-D intensity image. This is the first step for almost all the robust face recognition and analysis systems. Face models can be further simplified in different ways according to the analysis strategies, our interests and applications.

In the feature-based strategy, faces are modeled by a feature vector consisting of locations, distances or angles between feature points extracted from the front view faces or profile faces.

In template-based strategies, faces are modeled by a composition of grey scale templates, which represent significant facial regions, such as the nose or the eye area. The initial work for face templates can be found in [21]

Faces in the principle component analysis (PCA) approach are modeled by a set of weighting parameters for eigenfaces, which are linear combinations of sample variations

from the mean sample. Faces modeled in this way have various applications including face recognition, face compression and vector quantization.

The above is just a general description on face models. Many models are combinations of the above ones. If the distinct information provided by a face boundary is of interest, another 2-D model can be obtained. For example, in [20], the original face image is transformed to a binary one consisting only of edges by a special operator named *a valley operator*. The binary image is then used for visual communication where facial detailed texture is not important. Another example is shown in [19], where a flexible face shape model is constructed by coordinates of 152 landmarks representing main facial features and then used for face recognition.

3-D Face Models

In this section, we will look at faces as 3-D structures. Three dimensional facial models have been proposed for use in videophones where scenes generally restricted to head-and-shoulder types must be transmitted over low-bandwidth channels.

From a priori knowdge about a human face, if we can construct the face model that leads to a 2-D facial image sequence, then facial image information can be represented by analyzing and synthesizing faces based on the model. Since only the analysis information needs to be sent, this type of image coding can realize image transmission at a very low bit rate with a high image quality (typical bit rate is from 500 bits/s to 1000 bits/s). This is the basic idea behind facial model-based coding.

Generally, an explicit 3-D face model is employed in this application. The advantage of such models is that 2-D images can be more accurately described and more easily manipulated. Another advantage is that hidden surfaces or occlusions can be

removed from 2-D projected images with a 3-D model, which is very difficult when using only 2-D models.

The block diagram of the model-based coding scheme for facial image is shown in Fig.2.6. The assumption is that the transmitter and the receiver both possess the same 3-D facial model and texture image at the beginning of the visual communication. During the process, the transmitter will analyse the input frame to extract the facial motion parameters and updating parameters. Then, the receiver will synthesize the output frame using these parameters.

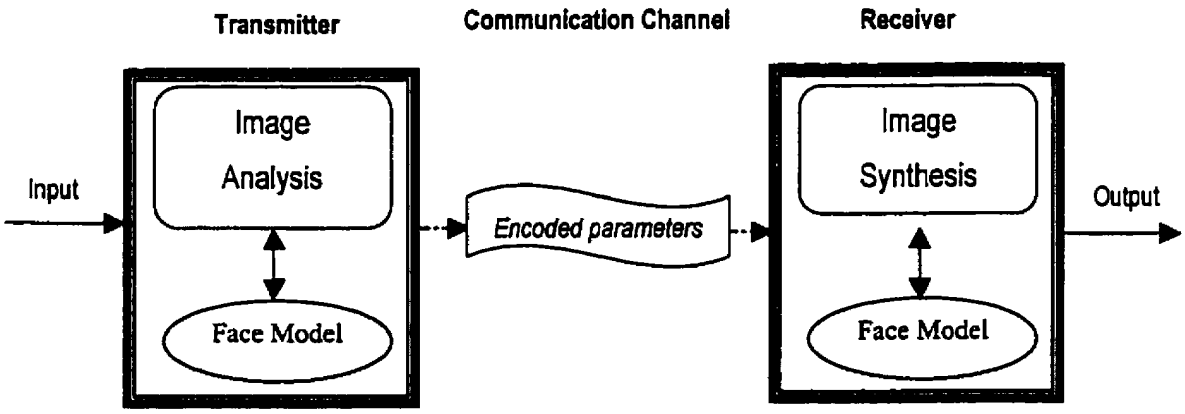


Fig.2.6 Block Diagram for Model-Based Coding

The 3-D models are always designed in a form of wireframe. According to whether anatomical information is employed or not during the wireframe deformation, the facial models can be classified to two types:

1. Geometric Model

The wireframe in this model describes the purely geometric appearance of the face and no anatomical information of the face is embedded. Normally, the wireframe model is composed by a set of connected triangular planar patches, where small patches are for high-curvature areas and larger ones for low-curvature areas. Examples can be derived from Parke's work in [7][8].

2. Physical Model

In this model, the face is represented as a lattice of point masses connected by elastic springs. The anatomical information on the movement of facial muscle groups is used and various facial expressions are generated through different combinations of facial tissues. An example is shown in [16].

In practice, the combination of both models is usually utilized to produce a realistic animated image of the speaker's face. An extension of this work is shown in [18], where the geometric wireframe model possesses physical information by mechanical constraints overlaid on the face motions or deformations.

Chapter 3 System Design

3.1 System Diagram

As mentioned in section 2.1.3, when several video frames of a static scene are available, noise reduction can be accomplished simply by averaging all the frames. This process, the simplest temporal filter called *frame averaging*, is also the basic idea behind the enhancement approach in this thesis.

However, movement of the objects (or the camera) often prevents this in practice. For example, a security video will probably provide a frame sequence of an individual in motion, where the size, orientation and expression of the person's face changes constantly. Our strategy for enhancing facial images is to employ *warping* to eliminate these variations and to produce a uniform image sequence which will be involved in frame averaging.

Fig.3.1 is the block diagram for the system. Suppose we first get an original sequence of moving facial images captured by a video camera. After some preprocessing steps, all frames in the sequence are warped to a selected frame (we call it *the target frame*). Then a new sequence in a uniform shape (we call it *the target shape*, which is subject to the target frame) is generated and frame averaging is performed on all deformed frames in this new sequence. The averaging result will be *the recovered target frame*, which can be used for identification. Finally, the desire to enhance the full sequence can be satisfied by warping the recovered target frame to each frame in the original sequence.

The preprocessing stage of the system includes histogram equalization and shade removal. For very dark or very bright images, histogram equalization, which has been discussed in section 2.1.1, will be performed to enhance the contrast. For images taken under nonuniform lighting conditions, a shade removal technique will be used to balance the shadings in the face area. The specific scheme for shade removal will be described at the end of this chapter in section 3.4.

Warping is applied in the enhancement process by overlaying a 2-D face model on the input frames. Each landmark in the face model serves as the control point during transformation. Section 3.2 provides the definition of this 2-D face model and corresponding landmarks, while section 3.3 deals particularly with the warping algorithm.

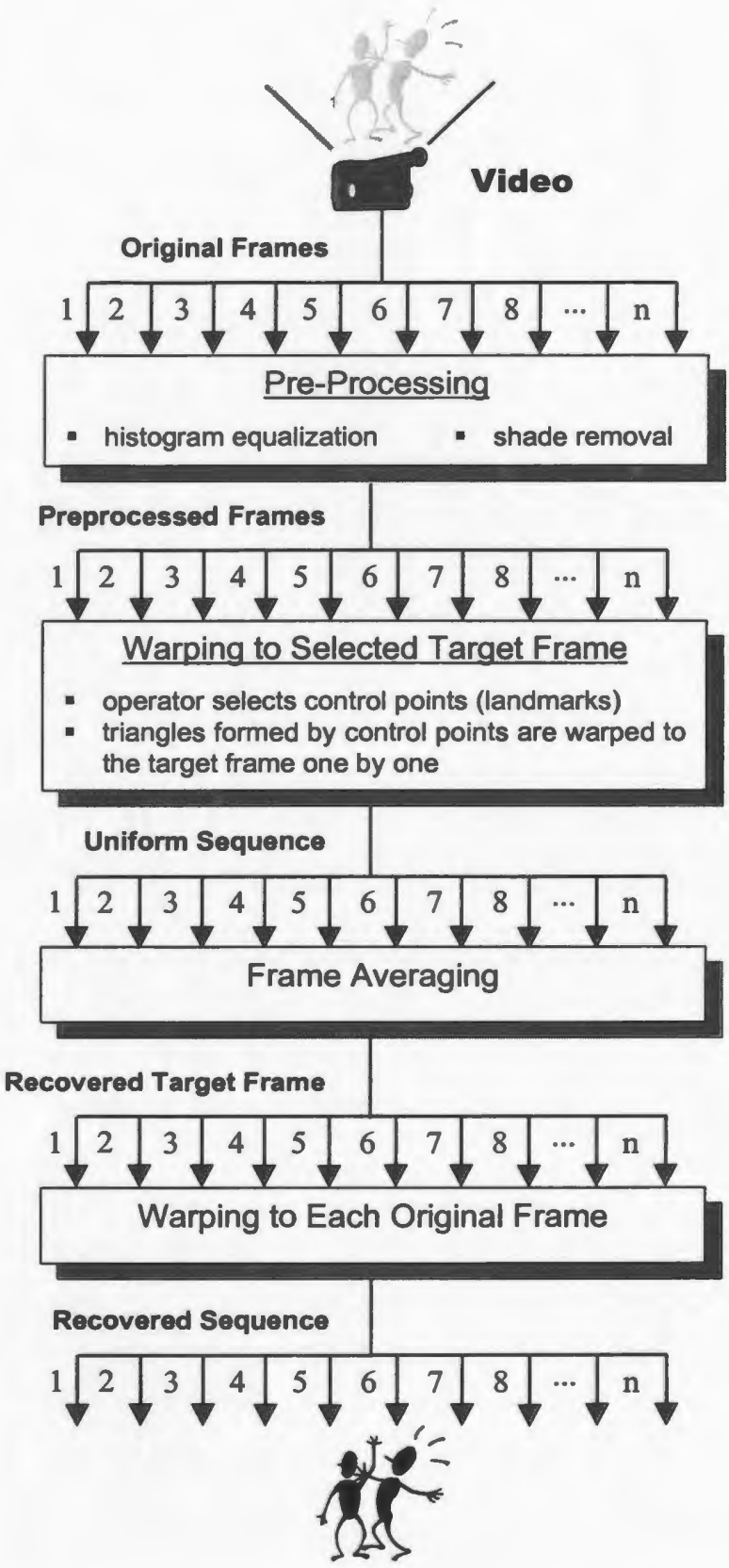


Fig. 3.1 Diagram of the Enhancement System

3.2 Landmarks and 2-D Face Model

Landmarks are the representations of the main features on a facial image. The coordinates of the landmarks constitute the 2-D face model in this thesis and provide clues for face warping.

How to select appropriate landmarks to describe the face layout accurately and efficiently is an important issue. In our work, the principles for picking the landmarks are as follows:

- landmarks should represent distinctive features of human faces.
- since the identification is made manually, landmarks should be easily located by eye, i.e., the grey level intensity should be quite different from its surroundings.

As mentioned in section 2.2.2, 152 landmarks are used in [19]. To balance the following factors, where the first two would tend to more landmarks to avoid artificial effects and the others less to save work, I determined empirically that 33 landmarks are required for the face model in the specific situations of this thesis:

- to have a realistic face appearance
- to accommodate the range of face shapes and expressions
- to identify them manually
- to identify them in images of poor quality
- to warp faces efficiently

The landmarks can be divided into three types:

Type 1: extremities on the outline of the face boundary and major organs on the face such as eyes, nose and mouth.

Type 2: inflection points on the outline of the face boundary and major organs on the face such as eyes, nose and mouth.

Type 3: points which can be calculated from the above two types.

In Fig.3.2, a face model with landmarks overlaid on it is shown. The detailed definitions of these landmarks are also attached. In this 2-D face model, the subsets of landmarks depict the shape of the main facial organs. In particular, landmarks 3 to 7 and landmarks 8 to 12 are for the left eye and the right eye respectively, landmarks 22 to 30 are for the face boundary, landmarks 16 to 21 are for the basic shape of the mouth, landmarks 1 to 2 indicate the eyebrow locations, and landmarks 13 to 15 are for the nose (see Figure 3.2). It should be pointed out that the landmarks numbered from 31 to 33 are obtained by calculation. The significance of these landmarks will be explained later. These landmarks altogether describe a realistic face appearance, including the information on expression, orientation and size of the face image. In addition, being significant facial features, they can be identified easily even in a poor quality image and only minor differences exist due to subjective adjudgements. Therefore, this 2-D face model consisting of 33 landmarks is applicable in the situations in this thesis.

As described later, each noisy frame will fit our face model by identifying all the corresponding landmarks. In theory, the frames partially occluded with a hand, a coffee cup, etc. so that some landmarks can not be located should be taken out from the system. The system relies on the frames with all the landmarks identified. This is the condition for the proposed system. However, for frames where only one or two landmarks are occluded by hair or nose, it is possible to locate these landmarks based on subjective estimations.

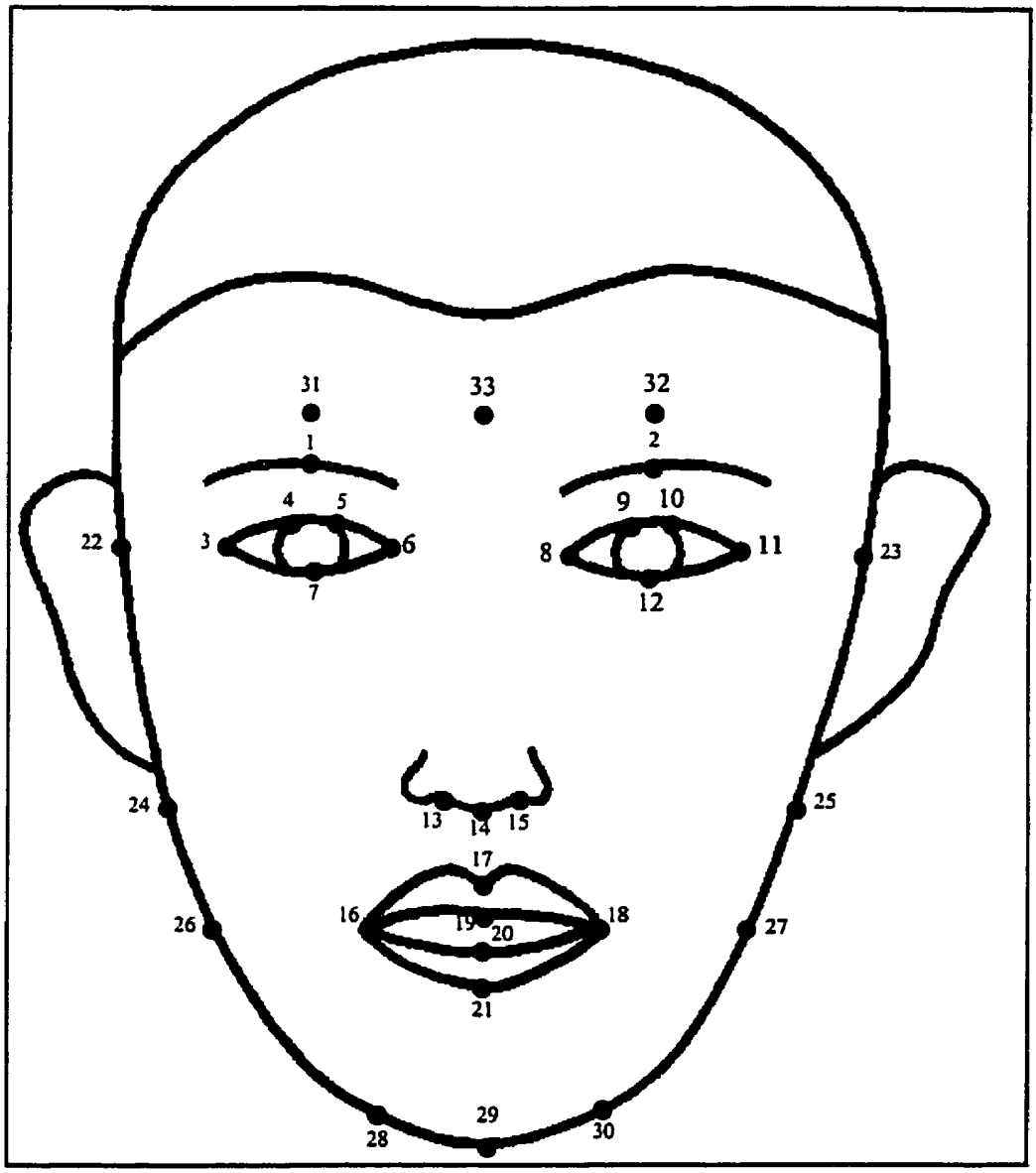


Fig.3.2 Definition of Landmarks

Outline of Eyebrows

1 – middle of upper contour of left eyebrow
2 – middle of upper contour of right eyebrow

Outline of Eyes

3 – outer corner of left eye
4, 5 – crossing points of eyeball with upper eyeliner
6 – inner corner of left eye
7 – left eyeball bottom
8 – inner corner of right eye
9, 10 – crossing points of eyeball with upper eyeliner
11 – outer corner of right eye
12 – right eyeball bottom

Outline of Nose

13, 15 – nostrils
14 – nose tip

Outline of Mouth

16 – left corner of lip
18 – right corner of lip
17 – middle of 16 and 18 on top of upper lip
19 – middle of 16 and 18 on bottom of upper lip
20 – middle of 16 and 18 on top of lower lip
21 – middle of 16 and 18 on bottom of lower lip

Outline of faces

22, 23 – eye line
24, 25 – nose line
26, 27 – mouth line
29 – bottom of face outline
28 – straight bottom of the left corner of mouth on the face outline
30 – straight bottom of the right corner of mouth on the face outline

Others

31 – five pixels row up than 1 and in the same column as 1
32 – five pixels row up than 2 and in the same column as 2
33 – midpoint of 31 and 32

Fig.3.2 Definition of landmarks (continued)

3.3 Warping

3.3.1 Background

Image warping is seen as a growing branch of image processing as it takes an input image and applies a geometric transformation to produce an output image. A geometric transformation is an operation that redefines the spatial relationship between pixels in an image by a mapping function.

Early interest in this area dates back to the mid-1960s in remote sensing, where they tried to reduce distortion. Today, it is also actively used in fields such as medical imaging, computer vision, and computer graphics [24].

Because a warp is a distorted view of an image or part of it, facial images can be deformed entirely or partially for caricatures or for morphing with a cross-dissolve of other warps [22]. Warping is used on moving facial images to produce a uniform sequence in this thesis.

3.3.2 Triangular Warping

Geometrical information, which can be obtained by triangles, rectangles, or lines, needs to be specified for warping. In our experiments, *triangular warping*, which corresponds to the affine transformation, is made on the original moving facial frames to generate a sequence in a uniform shape, i.e., the input face is divided into many triangles with landmarks in our face model (see Fig.3.2) as vertices. The deformations of the face shape in the aspects such as size, orientation and expression are based on this set of triangles.

The essentials for dividing the input face image into triangles are:

- to avoid triangles which are narrow and long.
- to avoid intersection of triangles due to movements of landmarks when orientation and expression are modified.

Fig.3.3 shows the triangulation scheme on a face layout explored in the thesis. Because the division is systematic, only the left side of the face is shown to make the figure easily readable. As mentioned before, landmarks numbered 31 to 33 are used to avoid long and narrow triangles.

There are 52 triangles altogether created by these 33 landmarks in the face model. The triangles cover the major area of the face from eyebrow to the lower jaw, which is a polygon composed by the outer landmarks. Hair and ears, which normally are not used to distinguish people, are removed from the face area.

Triangles are warped one by one, until the input frame is deformed to the shape of the target frame. Variation in the input frames is subject to the prerequisite that landmarks can be located.

3.3.3 Transformation Scheme

In triangular warping, each triangle on the source image is transformed to another one on the destination image. Landmarks, which are obtained by fitting the input and output frame to the 2-D face model, provide a straightforward mapping from triangle vertices on the input frame to specific locations on the target frame. Hence, the landmarks, also as triangle vertices, can be used as control points to calculate the transformation matrix T between the triangles.

Let (u_i, v_i) and (x_i, y_i) for $i = 1, 2, 3$ be these three landmarks in the source and destination triangles respectively (see Fig.3.4). The transformation matrix denoted as **T** is then simply [24]:

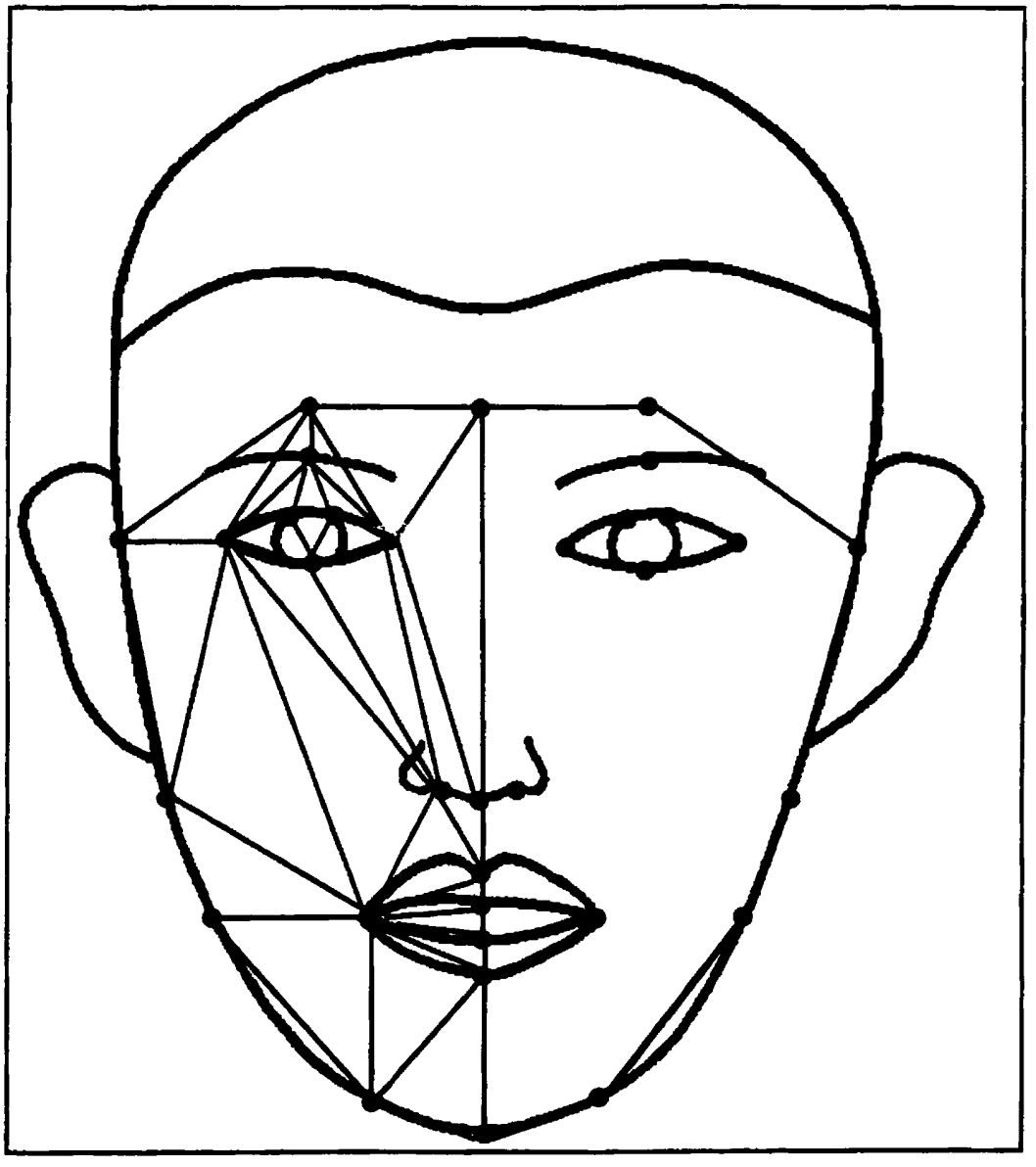


Fig.3.3 Triangular Division

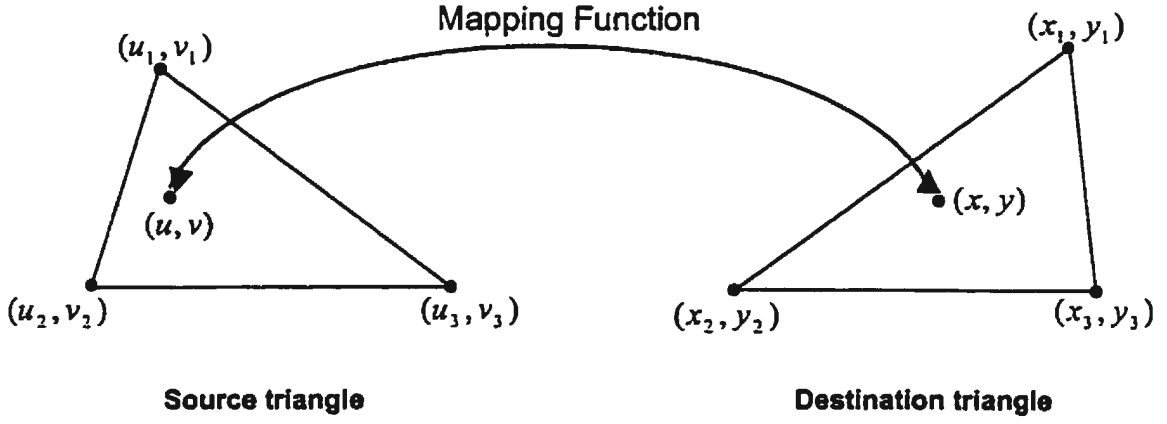


Fig.3.4 Triangle Mapping

$$\mathbf{T} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \tag{3.1}$$

Assume that $[u, v]$ and $[x, y]$ represent pixels within the source triangle and destination triangle respectively (see Fig.3.4). Equation (3.2) shows the calculation to specify the pixel positions between the triangles.

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{T} \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3.2}$$

Equation (3.2) is written in the format of *reverse mapping*, which goes through each pixel in the destination triangle and samples an appropriate source image pixel. There is another method called *forward mapping*, in which each pixel in the source triangle is mapped to an appropriate place in the destination triangle. Obviously, some pixels in the destination triangle may not be mapped by this method. We employ reverse mapping to avoid "holes" in the destination triangle so that all destination triangle pixels are guaranteed to be computed.

Therefore, all triangles in the input frame will be deformed to the target shape in the way shown in equation (3.2). For each particular triangle, the transformation matrix is subject to the corresponding landmark group which determines the shape of the current triangle.

However, the problem remains to determine when a pixel in the output space maps to a position which is between the discrete pixels in the input space. If the grey level of one of its nearby pixels (for example the nearest pixel) is simply taken, jagged edges will be perceived easily in the warped result. To solve this problem, filtering is necessary to integrate the grey level intensities surrounding the projected position. This process is called *interpolation*. To allow interpolation to occur in the input space instead of the output space is another advantage for inverse mapping. This is obviously a much more convenient approach than forward mapping.

For any pixels in the output space, the exact projected position P(X, Y) in the input space with respect to its surrounding existing pixels is shown in Fig.3.5. Our algorithm for interpolation is to give different weights to its nearby four pixels, i.e., the grey level of P(X, Y) is the sum of its surrounding 4 pixels by different weighting functions (see Fig.3.5).

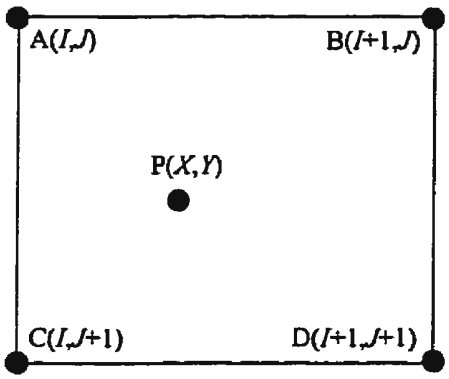


Fig. 3.5 Interpolation

Suppose $I < X < I + 1$ and $J < Y < J + 1$, then

- the weight of pixel A(I, J) is $[1.0 - (X - I)]\ [1.0 - (Y - J)]$,
- the weight of pixel B($I + 1$, J) is $[(X - I)]\ [1.0 - (Y - J)]$,
- the weight of pixel C(I, $J + 1$) is $[1.0 - (X - I)]\ [(Y - J)]$, and
- the weight of pixel D($I + 1$, $J + 1$) is $[(X - I)]\ [(Y - J)]$.

Such a choice of weighting functions is reasonable because the weight for a nearby pixel is proportional to its distance to P(X, Y). In addition, when P(X, Y) is coincident with one of the four surrounding pixels, the grey level is determined by that pixel since the weight for it is 1 while the weights for others are 0. Plus, when the surrounding pixels have the same grey level, any pixels within the boundary formed by the four pixels have a constant grey level since the sum of the weighting functions is 1.

$$[1.0 - (X - I)]\ [1.0 - (Y - J)] + [(X - I)]\ [1.0 - (Y - J)] +$$

$$[1.0 - (X - I)]\ [(Y - J)] + [(X - I)]\ [(Y - J)] = 1$$

This method is called *bilinear interpolation* [24]. Fig.3.6 shows two images obtained with and without interpolation.

without interpolation with interpolation

Fig.3.6 Effect of Interpolation

3.4 Shade Removal on Facial Images

The quality of facial images may be worsened if they are taken under an uncontrolled lighting conditions because of the 3-D structure of human faces. For example, different light source locations can cast or remove significant shadows from a particular face, hence bringing about the difficulty of face recognition. Therefore, shade removal techniques should be considered for frames in such a situation.

The strength of the heavy shadows caused by extreme lighting angles can be alleviated by the simple algorithm described in the following steps:

Step 1: a brightness surface from a facial frame is generated.

Step 2: the brightness surface is subtracted from the original facial frame.

Step 3: after surface subtraction in the second step, the grey level of the enhanced image is always increased or shifted. Hence grey level re-scaling is done to fit the grey level into the original range which is [0,255] in our case.

In this section, the construction of the brightness surface in *step 1* will be investigated. In addition, some pictorial results will be displayed.

3.4.1 Composing the Brightness Surface

To construct a surface that simulates the brightness over the complete image is the first step of the shade removal algorithm adopted in the thesis. The simplest surface adapted from [23] is a plane composed by two regression lines in the row direction and column direction respectively. The brightness plane constructed in this way is shown in Fig.3.7. In the *x*-*y*-*z* co-ordinate system, x indicates the row direction, y indicates the column direction, and z indicates the grey level intensity. The facial image lies in the *x*-*y* co-ordinate plane.

The two regression lines represented by linear equations $z = k_1 x + p_1$ and $z = k_2 y + p_2$ are derived from two sets of data: the average grey levels for all rows and for all columns on the facial frame. Using the "best-fit" method, the parameters for both lines can be calculated. The details of the calculation for the regression lines are introduced in *Appendix 1*.

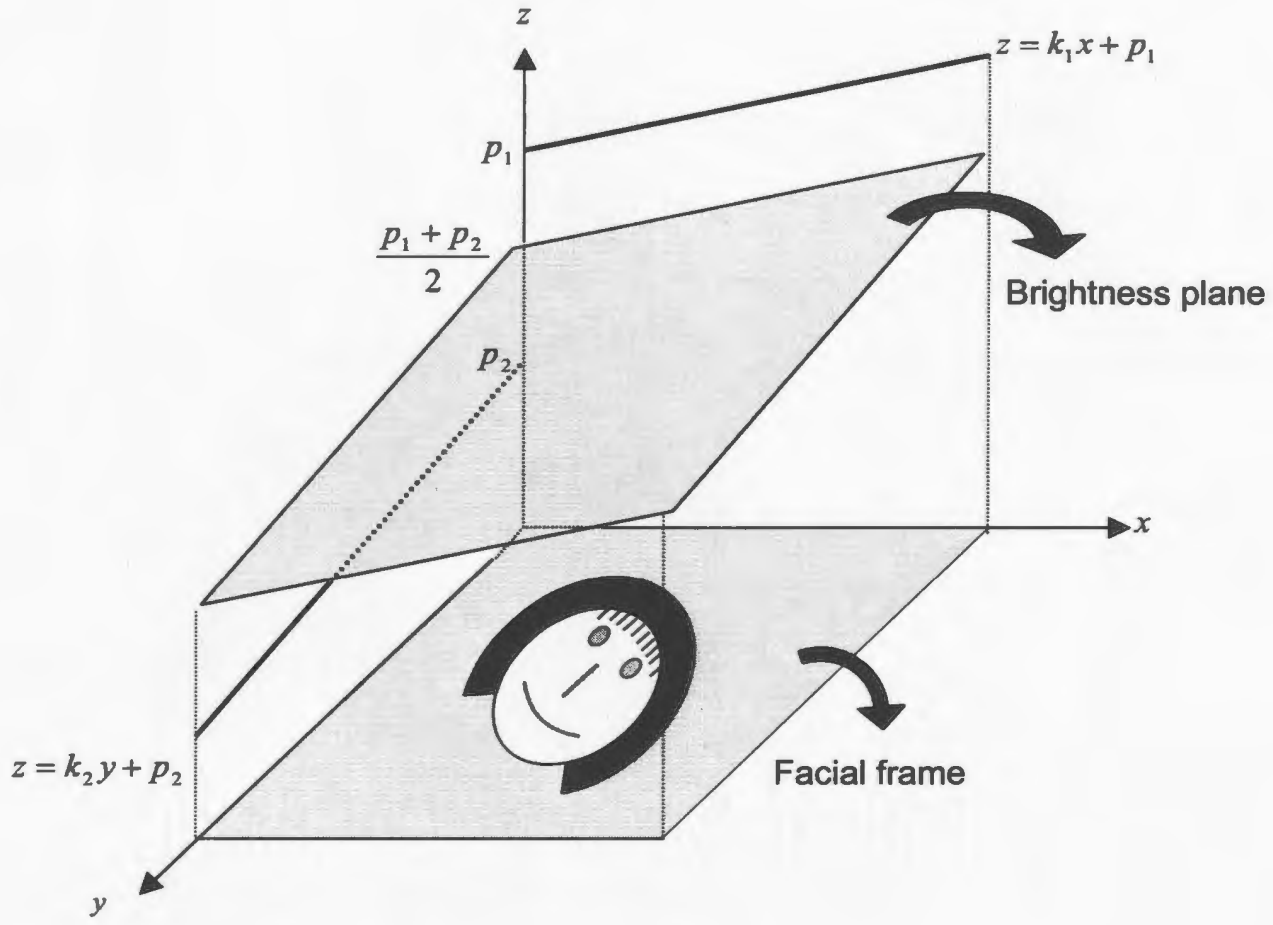


Fig.3.7 Construction of the Simplest Brightness Surface

Generally, a plane can not be generated directly from two regression lines because p_1 and p_2, the intercepts for the two lines, are always different at the z-axis. To compose the brightness plane, both lines are shifted in parallel and intersect at $(0,0,\frac{p_1 + p_2}{2})$, as shown in Fig.3.7. We will see the intercepts will not be considered in further discussion. *Appendix 2* describes the procedures in calculating the plane equation.

The weakness of using this brightness plane is apparent. First, our interest is in the *facial mask*, the region bounded by outer-most landmarks inside each frame. But pixels

outside the mask which may represent the background also affect the construction of the brightness surface. Second, as mentioned before, re-scaling over the whole image is used to narrow the grey level range. The side effect of re-scaling is the loss of facial details.

To solve the first problem, the algorithm is improved in this thesis by fitting the regression lines only from the data within the facial mask. Intensity values of the pixels outside the mask will be ignored in computing the lighting variation across the face area. Work is also done to restrict the dynamic range of the grey level intensity on the brightness surface which results in the next problem. This is accomplished by dividing the brightness surface into 9 continuous sections instead of extending the brightness plane for the facial window to the entire image, as depicted in Fig.3.8.

We define the minimum rectangle that contains the integrated facial mask as *the facial window*. On a facial frame, the region numbered 5 represents the facial window. The brightness surface for region 5 is constructed by two regression lines. The brightness surface for regions numbered 1, 3, 7 and 9 are individual planes parallel to the *x-y* co-ordinate plane, with intensity equal to the top-left corner, top-right corner, bottom-left corner and bottom-right corner of the brightness plane for region 5. The brightness surface for regions numbered 2, 4, 6 and 8 are determined by connecting their neighbouring planes. Please refer to Fig.3.9 for a 3-D view of the brightness surface constructed in this way.

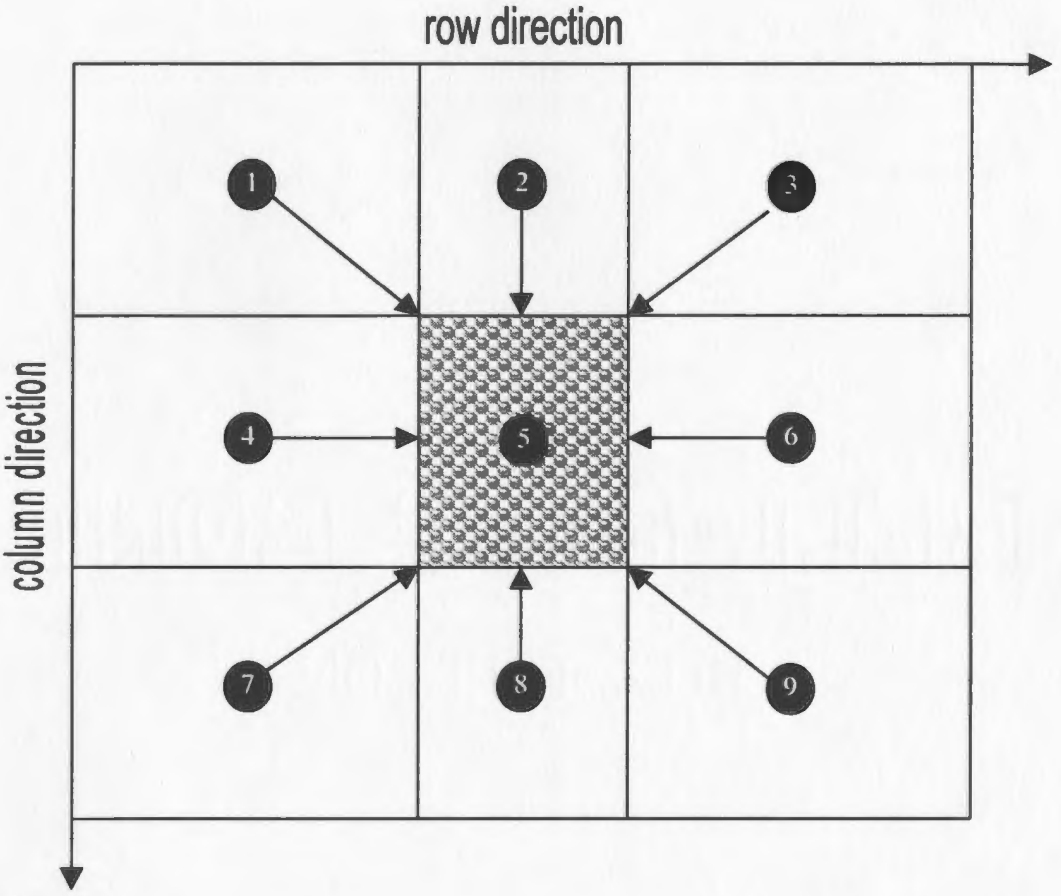
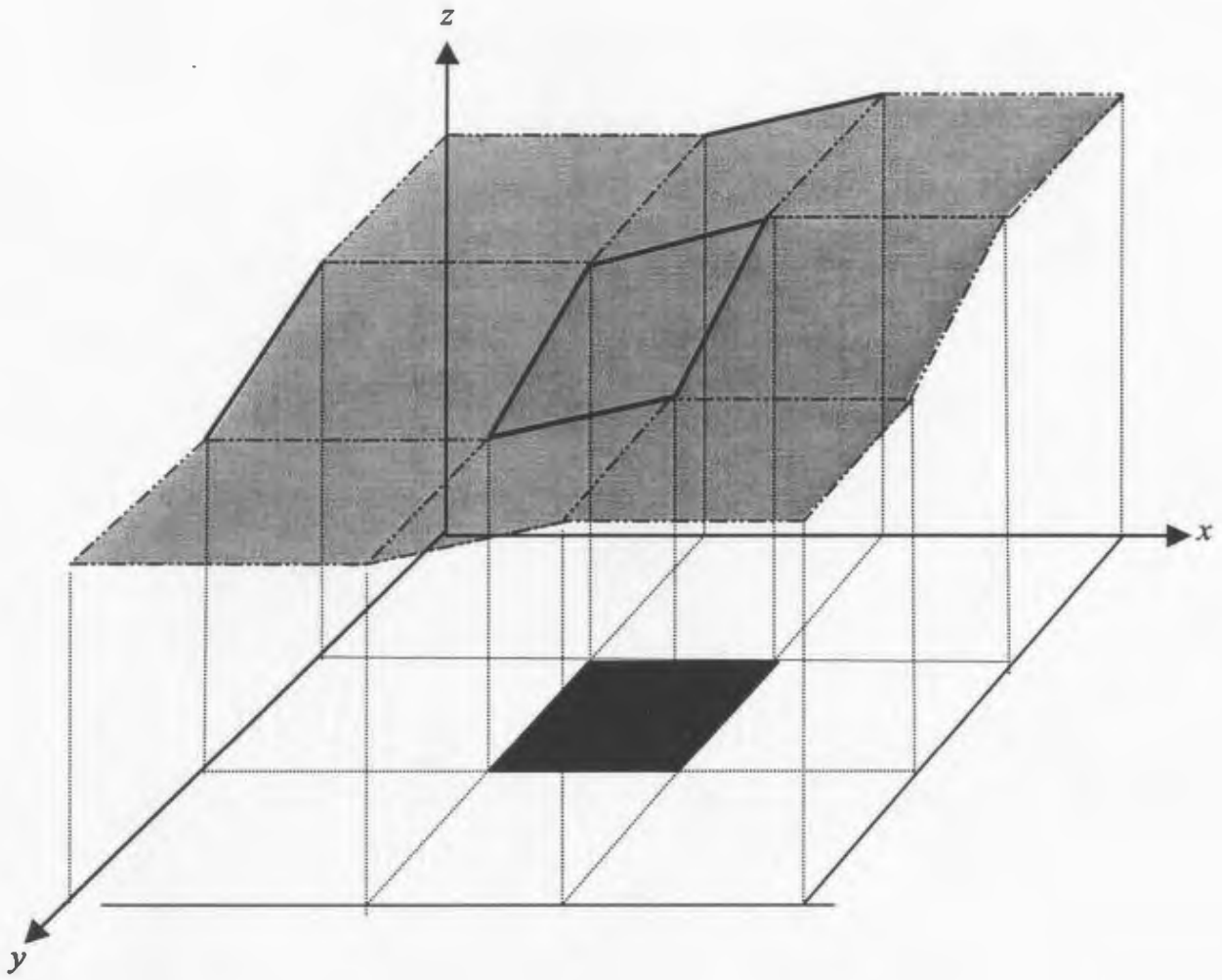


Fig.3.8 Division of the Brightness Surface



Brightness Surface

Facial Window

Fig.3.9 Enhanced Brightness Surface

It should be noted that only the slope of the regression lines are needed to construct the brightness plane for a facial window since we are only interested in the rate of change, which best fits the average trend curve. To compensate for a variety of lighting conditions while retaining the average grey level, the surface in Fig.3.9 is further modified in two steps: the first is to move the origin from (0, 0, 0) to the middle point of the facial window, the second is to shift the surface down vertically to pass through this new origin.

If [*x_min*, *y_min*], [*x_max*, *y_max*] and [*x_mid*, *y_mid*] represent the top-left corner, bottom-right corner and the middle point of the facial window respectively, the equations for intensities on the final constructed brightness surface (represented as *shade*[*x*, *y*] for the pixel at [*x*, *y*]) can be written as follows:

For [*x*, *y*] in region 1:

$$shade[x,y] = k_1 \times (x\_\min - x\_mid) + k_2 \times (y\_\min - y\_mid);$$

For [*x*, *y*] in region 2:

$$shade[x,y] = k_1 \times (x - x\_mid) + k_2 \times (y\_\min - y\_mid);$$

For [*x*, *y*] in region 3:

$$shade[x,y] = k_1 \times (x\_\max - x\_mid) + k_2 \times (y\_\min - y\_mid);$$

For [*x*, *y*] in region 4:

$$shade[x,y] = k_1 \times (x\_\min - x\_mid) + k_2 \times (y - y\_mid);$$

For [*x*, *y*] in region 5:

$$shade[x,y] = k_1 \times (x - x\_mid) + k_2 \times (y - y\_mid);$$

For [*x*, *y*] in region 6:

$$shade[x,y] = k_1 \times (x\_\max - x\_mid) + k_2 \times (y - y\_mid);$$

For $[x, y]$ in region 7:

$$shade[x, y] = k_1 \times (x\_\min - x\_mid) + k_2 \times (y\_\max - y\_mid);$$

For $[x, y]$ in region 8:

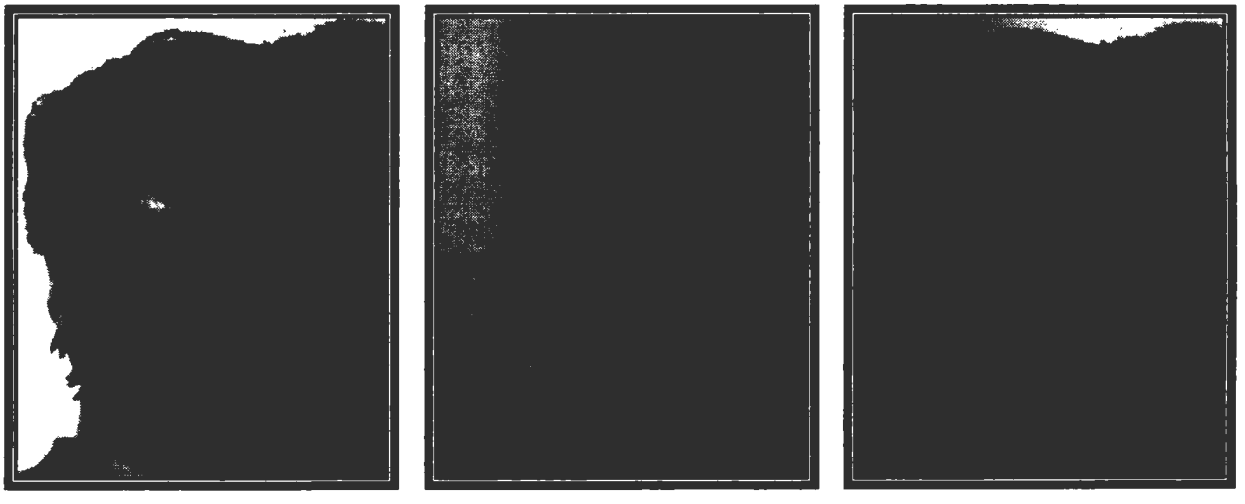
$$shade[x, y] = k_1 \times (x - x\_mid) + k_2 \times (y\_\max - y\_mid);$$

For $[x, y]$ in region 9:

$$shade[x, y] = k_1 \times (x\_\max - x\_mid) + k_2 \times (y\_\max - y\_mid);$$

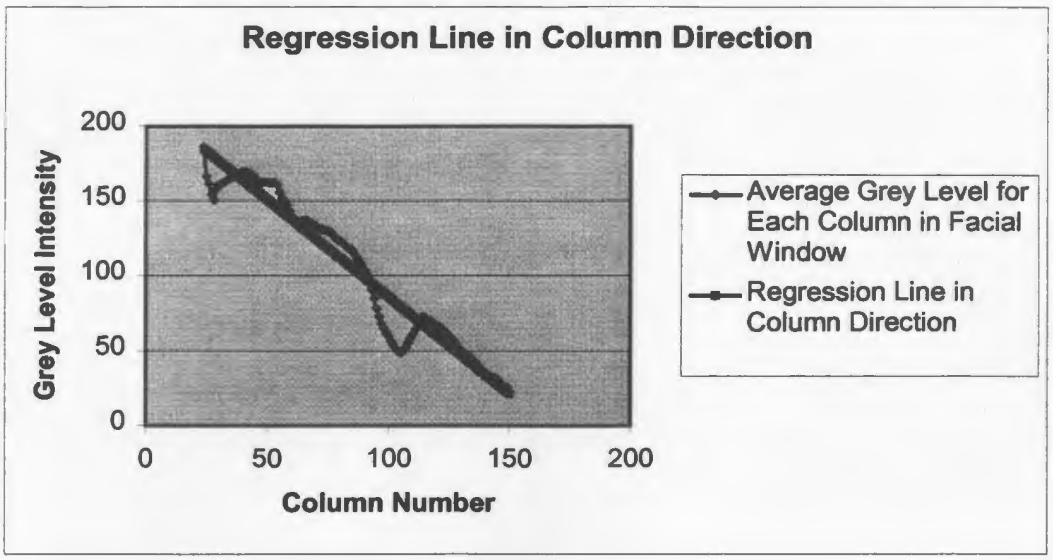
3.4.2 Pictorial Results for Shade Removal

A facial image in poor lighting conditions is shown in Fig.3.10(a). The brightness surface constructed in the way just described is shown in Fig.3.10(b). By subtracting the brightness surface from the original image, we get the shade removal result as in Fig.3.10(c). The dark area on the right side of the image can be clearly detected now. The shape of the regression lines in row and column directions are shown in Fig.3.11.



(a) original frame (b) brightness surface (c) after shade removal

Fig.3.10 Shade Removal Result



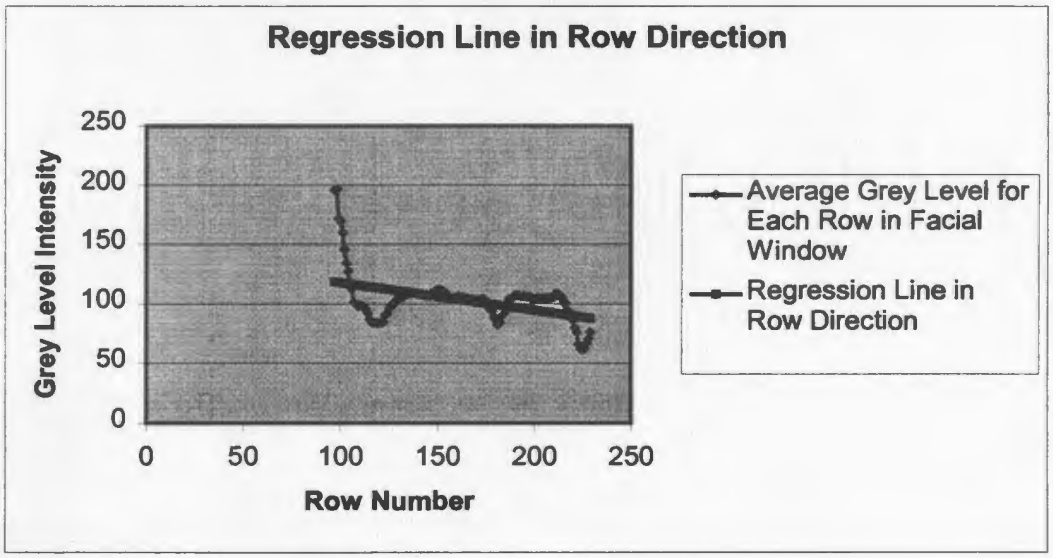


Fig.3.11 Regressions for Facial Image in Fig.3.10(a)

(a)

(b)

(c)

Fig.3.12 Other Removal Results

(d)

Fig.3.12 Other Results of Shade Removal (continued)

Chapter 4 System Implementation

4.1 Interface Description

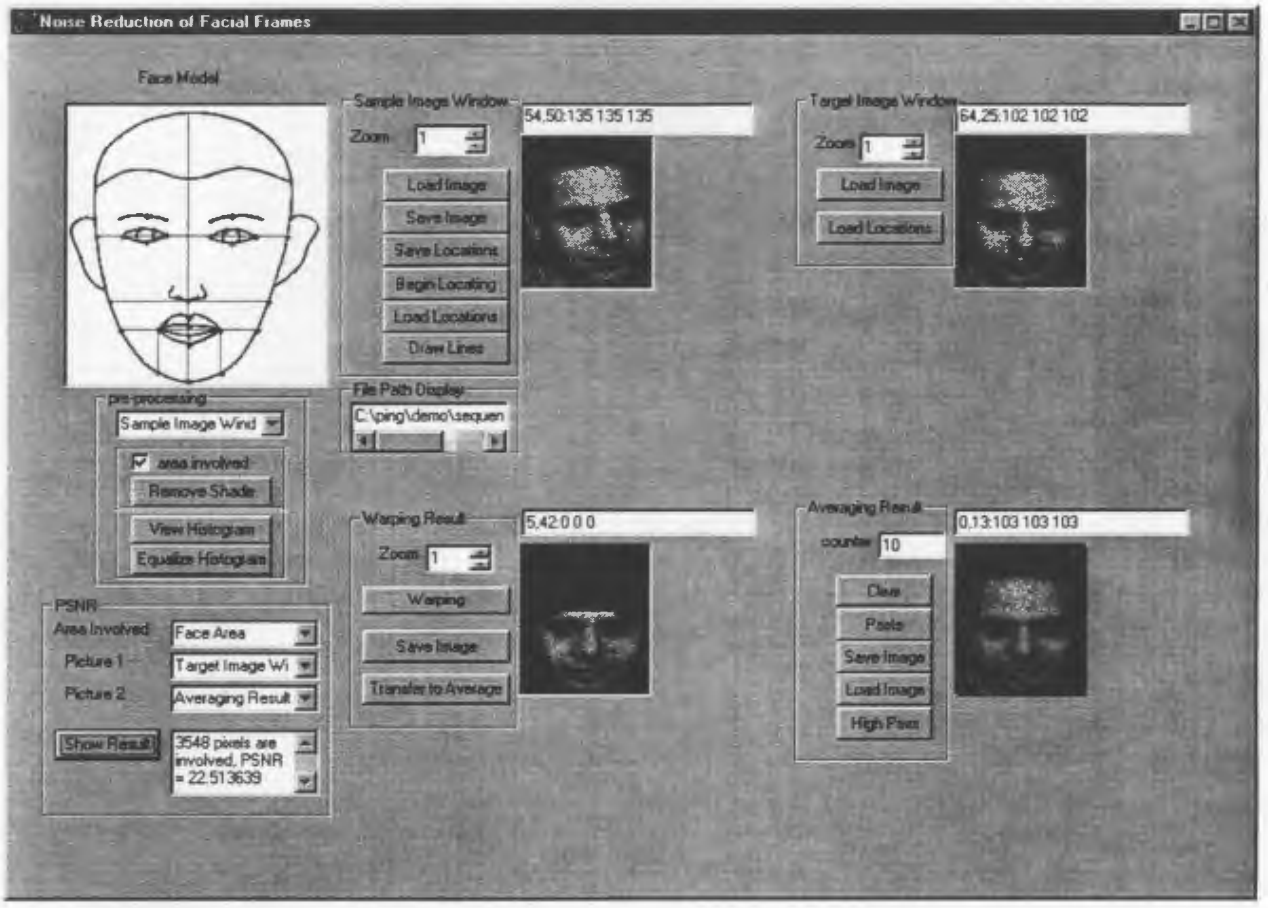


Fig. 4.1 Interface for Noise Reduction of Facial Frames

All the functionality described in Chapter 3 is provided in a Windows program with a simple user interface. The program is written in C++ and uses MCLGallery developed by Li-Te Cheng [25]. Fig.4.1 shows the interface applied to an example image.

With the aid of this interface, the face model can be displayed, landmarks located manually, all noisy frames transformed to the target shape by warping, and frame

averaging on the transformation results performed. The interface also has the facilities to preprocess the images and compare the difference between two images by PSNR calculation.

4.2 Interface Composition

The interface is composed by five windows which contain images: *Face Model*, *Sample Image window*, *Target Image Window*, *Warping Result*, *Averaging Result*, and two operation groups: *Preprocessing and PSNR*.

4.2.1 Face Model

The face model with landmarks represented by red dots is shown in the Face Model window. During the landmark locating process, each landmark will blink in turn to direct a proper locating sequence.

4.2.2 Sample Image Window

A facial image in JPEG, GIF, PGM or PPM format can be loaded into this window by clicking **`Load Image`** and corresponding landmarks can be overlaid on it by clicking **`Load Locations`** to select the appropriate data file. The updated image after preprocessing can be saved by clicking **`Save Image`**.

Landmark locating will be performed on a loaded image if **`Begin Locating`** is clicked. Because all the landmarks should be located in the appropriate sequence, the landmarks in the face model will blink in turn to indicate the next landmark to be located during the operating process. Each landmark is created by clicking the left button of the

mouse on the frame. An existing landmark in the wrong position can be remedied by re-editing when the right button is clicked on it. When all landmarks have been located, a message window indicating the end of the work appears. Thus all landmarks can be easily located in the appropriate sequence and positioned even by inexperienced users. The coordinates of the landmarks are stored in a data file used for warping.

There is no automatic checking of landmark consistencies. But a couple of operations can be applied to check the landmark sequence:

- Clicking **Draw Lines** causes the connection of proper landmarks to compose a face-like shape. But **Draw Lines** can not detect all sequence errors. For example, if we exchanged the sequence for the landmarks numbered 17 and 19 (see Fig.3.2), it would fail.
- Pressing **Shift** while clicking the left button of the mouse on the existing landmark causes the index of the landmark to appear in the textbox. For example, there must be something wrong in landmark sequence if the index for the nose tip is not 14 (see Fig.3.2).

4.2.3 Target Image Window

The target frame to which all others will be transformed can be loaded here by clicking **Load Image**. The landmark locations can be located by clicking **Load Locations**.

4.2.4 Warping Result

The command buttons for this window are **Warping**, **Transfer to Average** and **Save Image**.

The sample image loaded in the Sample Image Window will be transformed to the shape of the target image loaded in the Target Image Window according to landmark assignments by clicking **Warping**. The facial area of the warping result will be displayed in the Warping Result window automatically. The result can be saved and participate in frame averaging by clicking **Save Image** and **Transfer to Average** respectively.

4.2.5 Averaging Result

The number of frames involved in frame averaging is recorded in a variable **counter**.

Before frame averaging, the counter should be emptied and the Averaging Result window reset to a completely black image by clicking **Clear**.

At the end of frame averaging, the facial image of the averaging result can be pasted to the background of the image in Target Image Window to have the recovery result which can then be deblurred by clicking **High Pass**. Please see Section 2.1.2 for details on the selected high-pass filter. The final result can be saved by clicking **Save Image**.

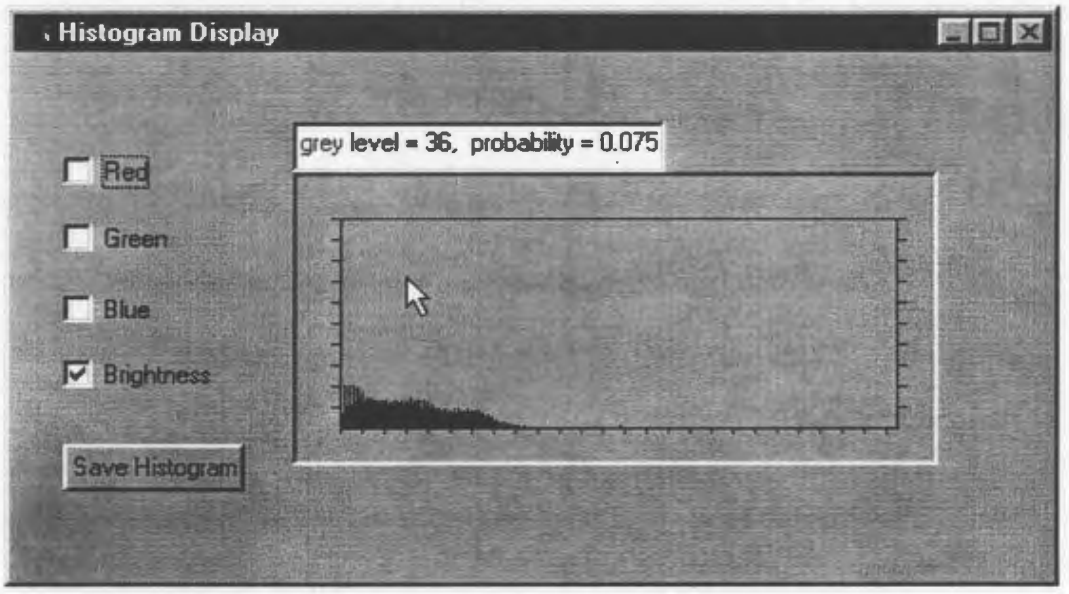
Load Image has the same function as in the other windows.

4.2.6 Preprocessing

Operations on histogram and shade removal technique are embedded in the Preprocessing group.

Operations on Histogram

By clicking on **`View Histogram`**, a Histogram Display window (see Fig.4.2) which shows the diagram of the histogram for the activated window will be displayed. If we move the mouse in the effective chart area, the grey level and the corresponding probability appeared in the image will be shown in the text box.



Note: The histogram displayed is for the original image in Fig.2.1(a)

Fig.4.2 Histogram Display Window

For color images, the histogram refers to frequencies of image color values, typically R, G, B values or brightness which is the average of R, G, B. In the Histogram Display window, the intensity of a color component, ranging from zero luminance (black) to full luminance (white), is indicated on the horizontal axis by checking the corresponding color channel. The portion of the image's color appears on the vertical axis.

By clicking **Equalize Histogram**, the histogram of selected image is smoothed out. The selected image will be transferred to an image with enhanced contrast. The general algorithm for histogram equalization is in Section 2.1.1.

For color images, only the histogram of the luminance channel will be equalized, to retain the original chromaticity. Therefore, the following three steps will be performed:

Step 1: Given an image where color is represented by R, G, B values, we first transform the color coordinate system and represent a color by Y, C_1, C_2 values, where Y is the luminance channel to be extracted, and C_1 and C_2 are chrominance channels. The transformation is made by [26]

$$\begin{bmatrix} Y \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Step 2: We equalize the luminance channel Y by the mapping function introduced in section 2.1.1 with chrominance channels C_1 and C_2 fixed.

Step. 3: The equalization result is obtained by inverse transform of the updated luminance channel and C_1 and C_2 channels to R, G, B values.

The inverse transformation is denoted by

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix}^{-1} \begin{bmatrix} Y \\ C_1 \\ C_2 \end{bmatrix}$$

Shade Removal

The shade removal technique explored in the thesis is described in Section 3.4.

By clicking the down arrow of combo box in the Preprocessing Window, the image window where shade removal will be performed is selected.

By checking **Area Involved**, shade removal is applied only to the facial area. The prerequisite for this is to have landmarks overlaid on the selected image.

4.2.7 PSNR

PSNR value is a qualitative measurement for the difference between two images of the same size. Before calculation, three components of PSNR must be determined. Two images are selected from combo box **Picture 1** and **Picture 2**. Combo box **Area Involved** decides whether PSNR is calculated on the whole image area or only the facial area bounded by landmarks in Target Image Window.

The PSNR calculation begins after **Show PSNR** is clicked. The PSNR value and number of pixels involved in calculation will be displayed in text box.

The formula to calculate PSNR is as follows:

$$PSNR = \sum_{(i,j)\in\Omega} \frac{(P_R[i,j]-P_R^{'}[i,j])^2+(P_G[i,j]-P_G^{'}[i,j])^2+(P_B[i,j]-P_B^{'}[i,j])^2}{3\times N_\Omega}$$

where Ω denotes the selected area (either the entire image or the face area) and N_Ω the total number of pixels in Ω. $P_R[i,j]$, $P_G[i,j]$ and $P_B[i,j]$ are the red, green and blue components for the grey level of the pixel positioned at $[i,j]$ within Ω in the first image respectively, while $P_R^{'}[i,j]$, $P_G^{'}[i,j]$ and $P_B^{'}[i,j]$ are the red, green and blue components for the grey level of the pixel positioned at $[i,j]$ in the second image respectively.

Chapter 5 Basic Experiments

The system diagram of the enhancement scheme proposed in this thesis is shown in Fig.3.1. In this chapter, the pictorial results for degradation and the recovered target frame in the basic experiments will be presented. Experiments for the recovery of entire sequences is the topic of Chapter 7.

5.1 Experiment Procedures

To test the effectiveness of the strategy, applications on noisy and blurred sequences are necessary. In the experiments, the first step is the degradation to a known extent on the original (blur-free and noise-free) frame sequences. In this way, practical videos can be simulated and quantitative measurement between frames obtained.

There are two kinds of image degradation, *blur* and *the composition of blur and noise*, which will be examined.

Fig.5.1 is the experiment diagram. The experiments can also be explained by the following procedures:

Step 1: to obtain degraded images by adding noise to the original frames or to the blurred original frames.

Step 2: to preprocess the degraded frames by histogram equalization or shade removal subject to the frame quality. For the basic experiments in this chapter, the preprocessing step will be ignored. It will be involved in experiments for subjective tests in Chapter 6.

Step 3: to fit the preprocessed/degraded images to the 2-D object model, i.e., to identify the landmarks manually and deform all degraded images to the target one by warping.

Step 4: to add all N deformed images together and get the average one as the recovery image.

Step 5: If blur is involved in degradation, the deblur filtering described at the end of Section 2.1.2 will be used to sharpen the recovery image obtained in *Step 4*.

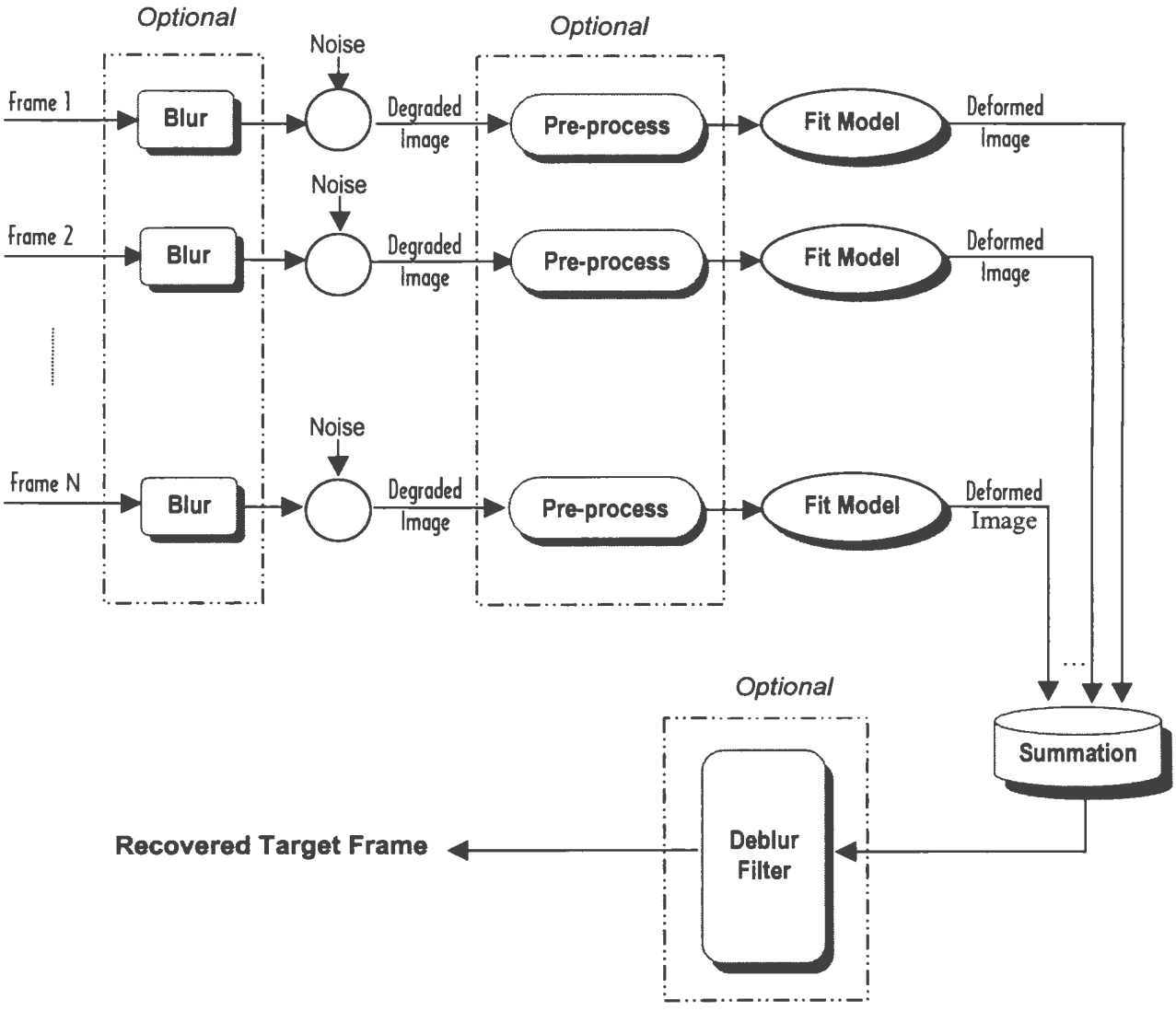


Fig.5.1 Experiment Diagram

5.2 Test Sequences and Degradations

The two original image sequences, which are taken from the Olivetti Research laboratory face database (http://www.cam-orl.co.uk/facedatabase.html) and composed of clearly discernible images, are shown in Fig.5.2. Each sequence features one person, termed the "candidate", and 10 frames (i.e., N=10) for each candidate with the same size of 92×112 pixels are used. The frames marked No.1 in Fig.5.2 will serve as the target frames for each candidate.

In the experiments, sequences with varying levels of added noise and blur degraded from these two sequences are tested. Specifically, degradation is implemented by adding noise directly or by the combination of blur and noise successively.

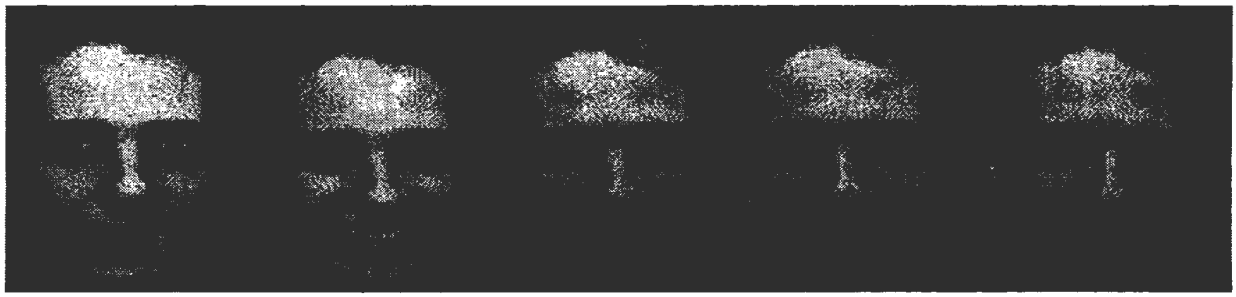
Noise is added by a program generating random integers according to the distribution of AWGN when the parameters of variance σ^2 and the mean value are given. Fig.5.3 shows the target frames for both candidates degraded only by noise with zero mean and variances of *0.001*, *0.005*, *0.01*, *0.03* and *0.05* (unit's scaling) respectively. When σ^2 is increased to *0.05*, the human faces are almost submerged by noise and can be hardly detected.

Blur degradation is accomplished by low-pass spatial filtering. The filter masks for slightly blurred, moderately blurred, and heavily blurred degradations, are provided in section 2.1.2. In practice, we usually acquire blurred frames with relatively little noise from a video sequence. The frames of Fig.5.4, in which the slightly blurred, moderately blurred and heavily blurred frames for each target candidate are destroyed by noise with variances σ^2=*0.001* and *0.005* respectively, are simulated for this situation.

No. 1 No. 2 No. 3 No. 4 No.5

No. 6 No. 7 No. 8 No. 9 No.10

(a) the First Candidate

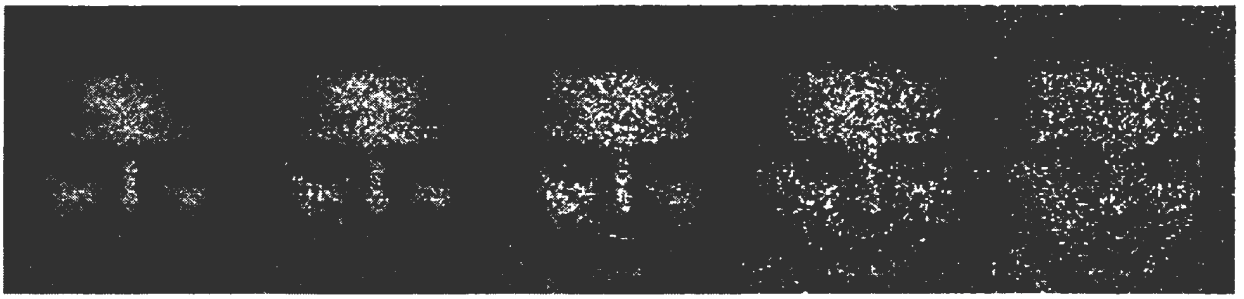


No. 1 No. 2 No. 3 No. 4 No.5

No. 6 No. 7 No. 8 No. 9 No.10

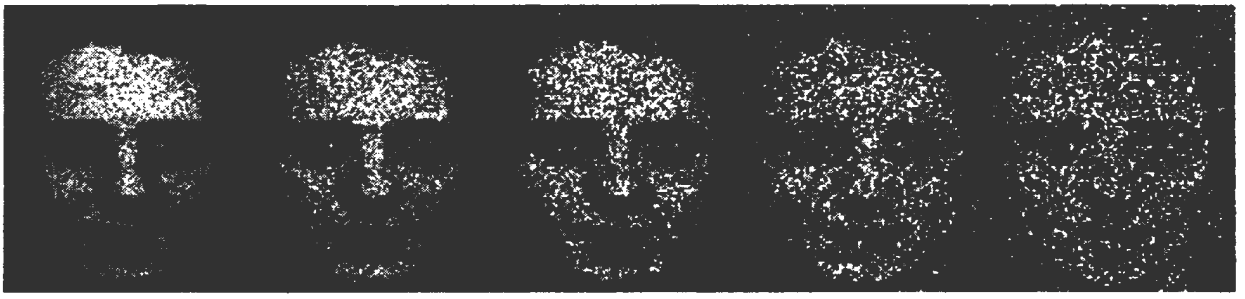
(b) the Second Candidate

Fig.5.2 Original Test Sequences



σ^2 =0.001 σ^2 =0.005 σ^2 =0.01 σ^2 =0.03 σ^2 =0.05

(a) the First Candidate



σ^2 =0.001 σ^2 =0.005 σ^2 =0.01 σ^2 =0.03 σ^2 =0.05

(b) the Second candidate

Fig.5.3 Noise Degradation of Target Frames

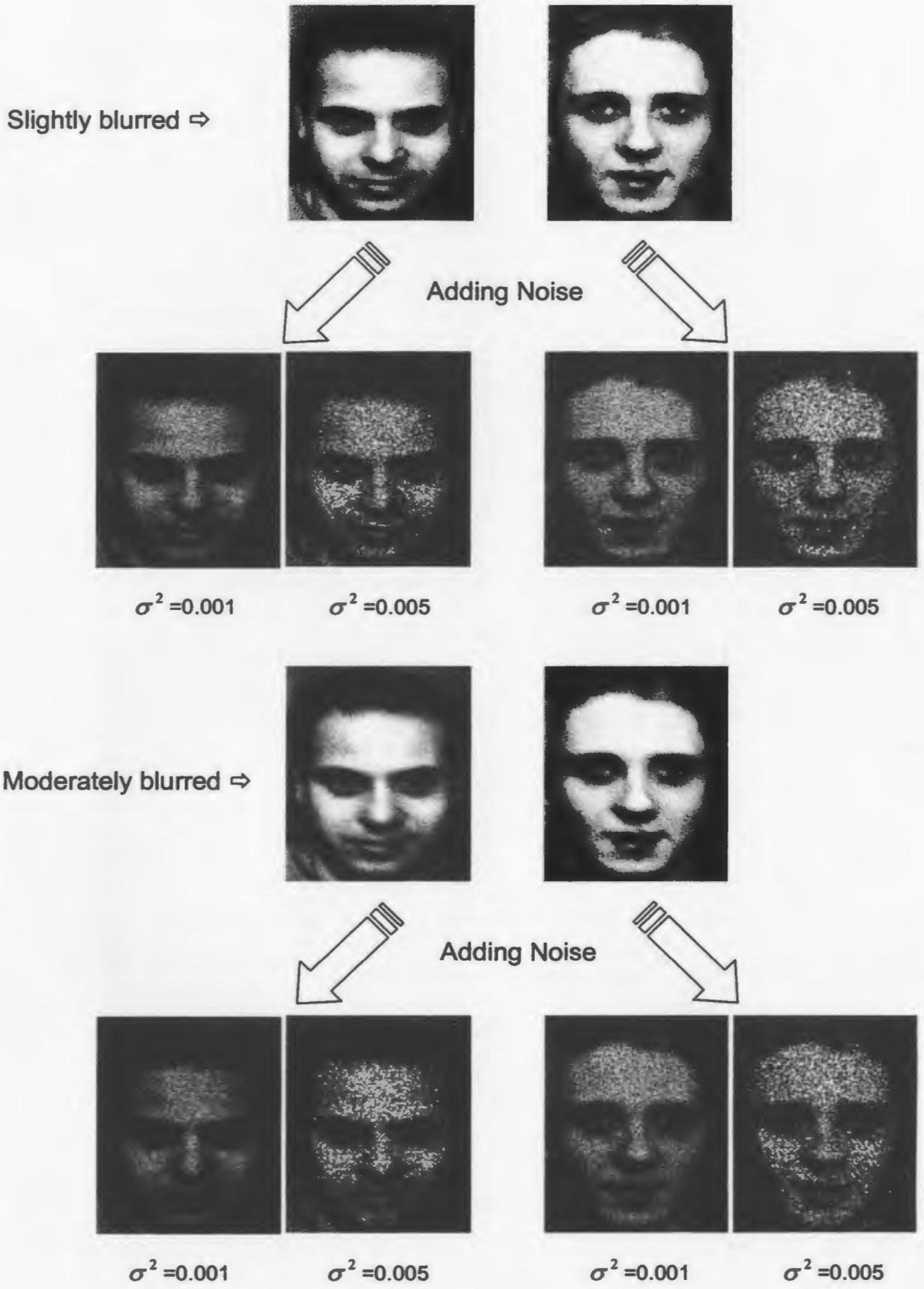


Fig.5.4 Blurring and Noise Degradations

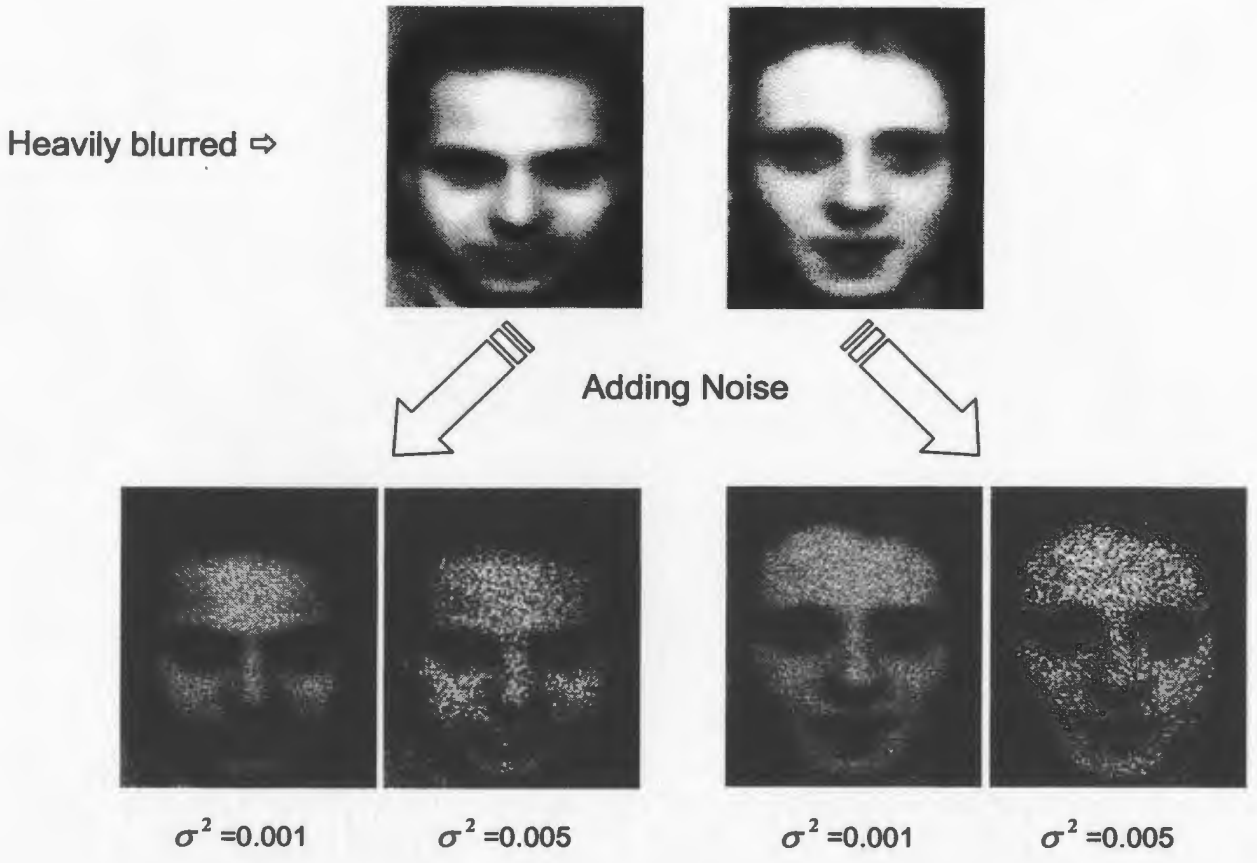


Fig.5.4 Blurring and Noise Degradations (continued)

5.3 Recovery from Original Images

In this experiment, landmarks are located on the original images shown in Fig.5.2. This leads to the optimum result of frame averaging after warping. The recovered frames for both candidates are shown in Fig.5.5. After comparing with the original target frame in Fig.5.2, it can be confirmed that the proposed recovery algorithm is feasible for perfect frame sequences. That is, it does not introduce processing artifacts. In the following experiments, the recovered target frame from various degraded sequences for the two candidates will be presented.

Candidate 1 Candidate 2

Fig.5.5 Recovery from Original Images

5.4 Recovery from Noisy Images

In this experiment, landmarks are located from the noisy images listed in Fig.5.3. Each degraded image of 10-frame sequences is deformed to the same target frame and then averaged as the recovered result, which is included in Fig.5.6. The corresponding variances are also attached.

The peak signal to noise ratio (PSNR) measured in dB for the noisy image before averaging and the recovered result obtained after averaging is listed in Table 5.1. More discussions of the issues around PSNR calculation are given in section 6.4.1.

5.5 Recovery from Noisy and Blurred Images

In this experiment, landmarks are located from noisy and blurred images as in Fig5.4. The final results obtained after deblur filtering are shown in Fig.5.7.

Note all recoveries (Fig.5.6 and Fig.5.7) can be further improved by introducing more frames into the averaging process.

σ^2 =0.001 σ^2 =0.005 σ^2 =0.01 σ^2 =0.03 σ^2 =0.05

(a) the First Candidate

σ^2 =0.001 σ^2 =0.005 σ^2 =0.01 σ^2 =0.03 σ^2 =0.05

(b) the Second Candidate

Fig.5.6 Recovered Target Frames from Noisy Images

Noise Variance σ^2	the First Candidate		the Second Candidate	
	Original (dB)	Recovered (dB)	Original (dB)	Recovered (dB)
0.001	30.2302	42.1986	30.1540	42.3874
0.005	23.1528	35.5455	22.9971	35.6553
0.01	19.7634	32.5278	19.3924	32.3878
0.03	14.6174	26.1368	14.1109	25.4543
0.05	12.1407	21.9942	11.8431	21.3385

Original PSNR:

Source image: degraded target frame (by noise).

Destination image: original target frame (noise-free and blur-free).

Involved area: facial mask obtained from degraded target frame.

Recovered PSNR:

Source image: recovered target frame from degraded frames (by noise).

Destination image: recovered target frame from original frames (noise-free and blur-free).

Involved area: facial mask obtained from degraded target frame.

Table 5.1. Comparison of Peak Signal to Noise Ratio (PSNR)

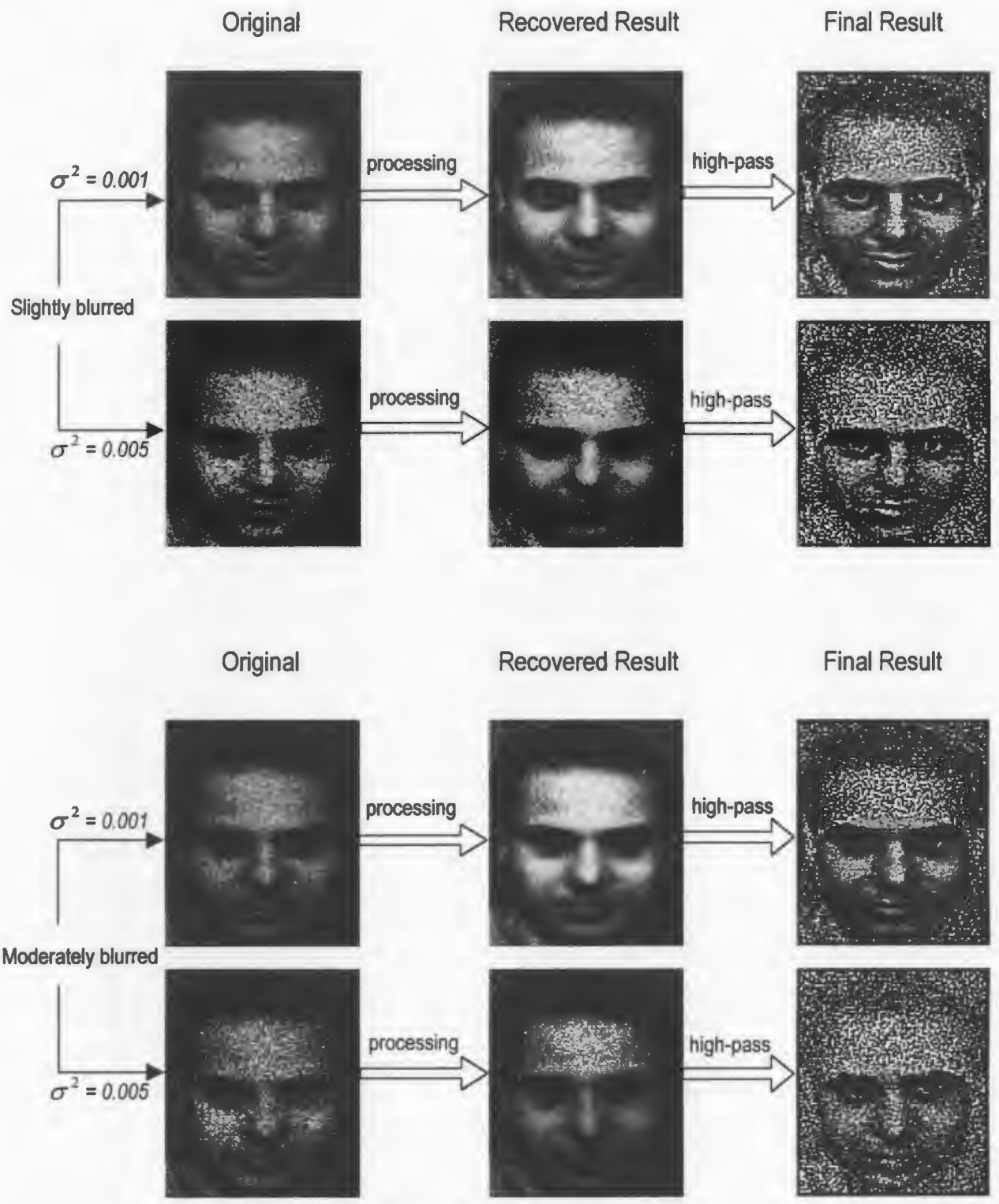
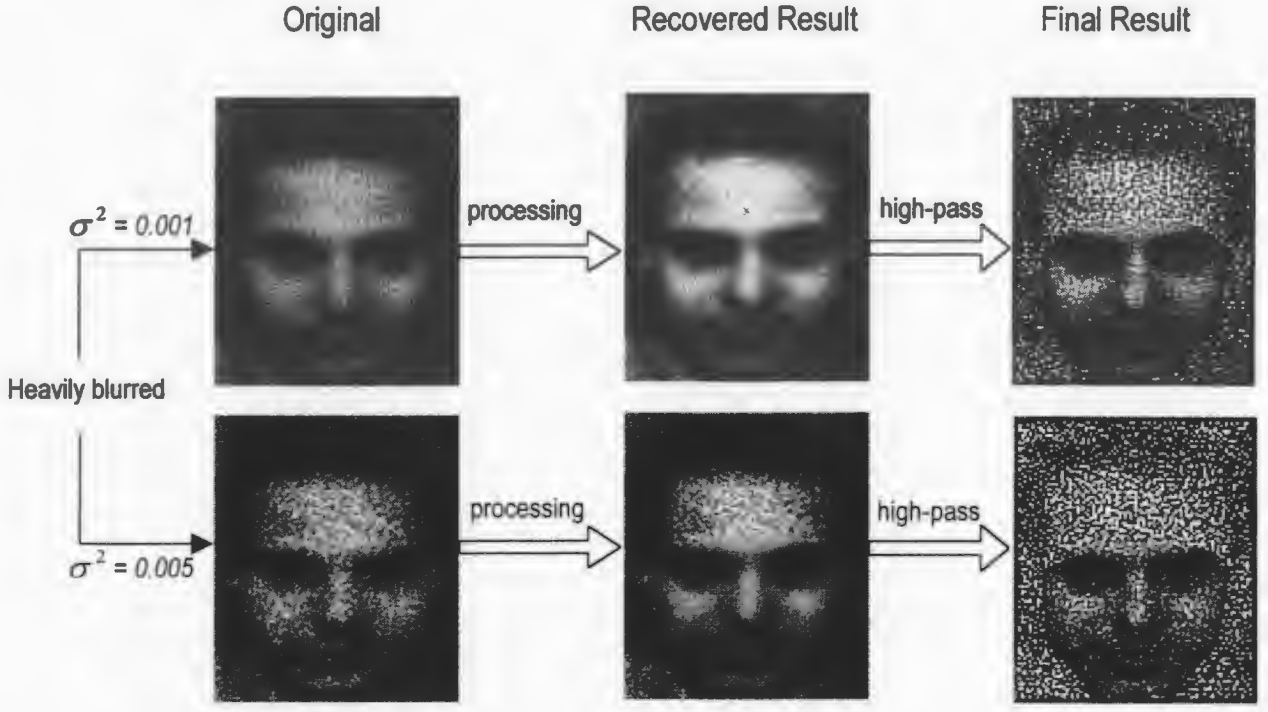


Fig.5.7 Recovered Target Frames from Noisy and Blurred Images



(a) the First Candidate

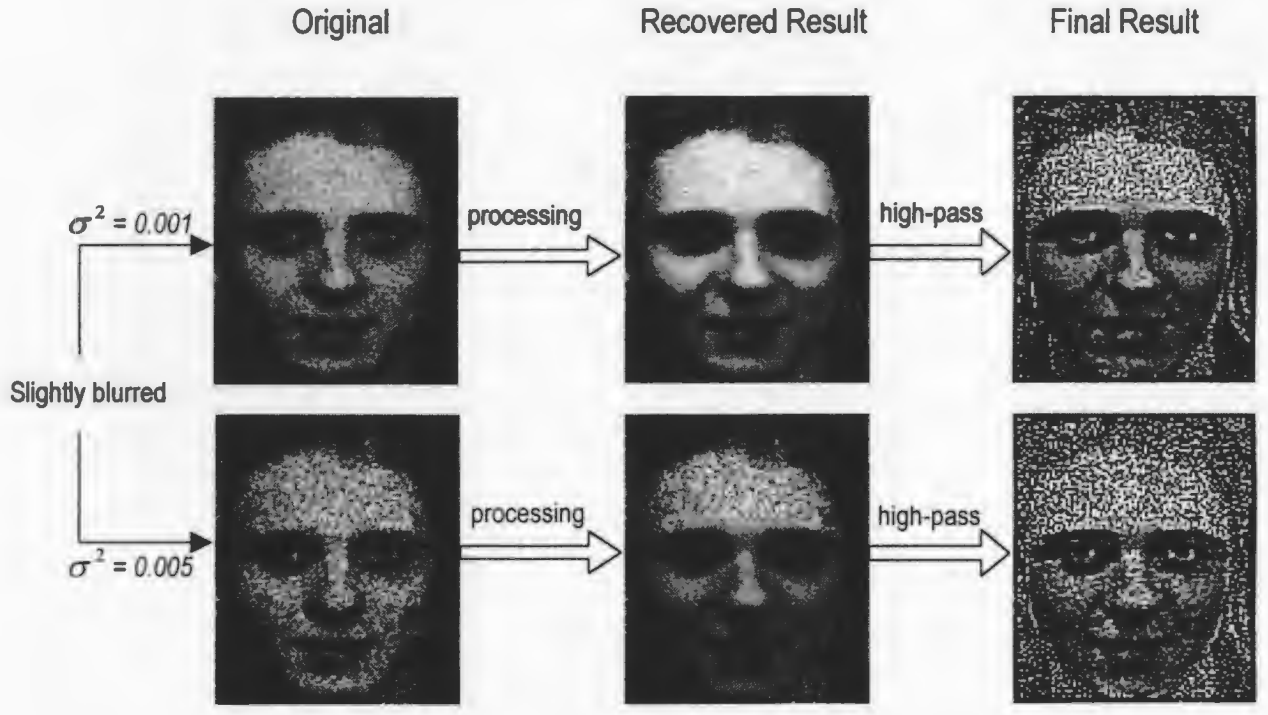
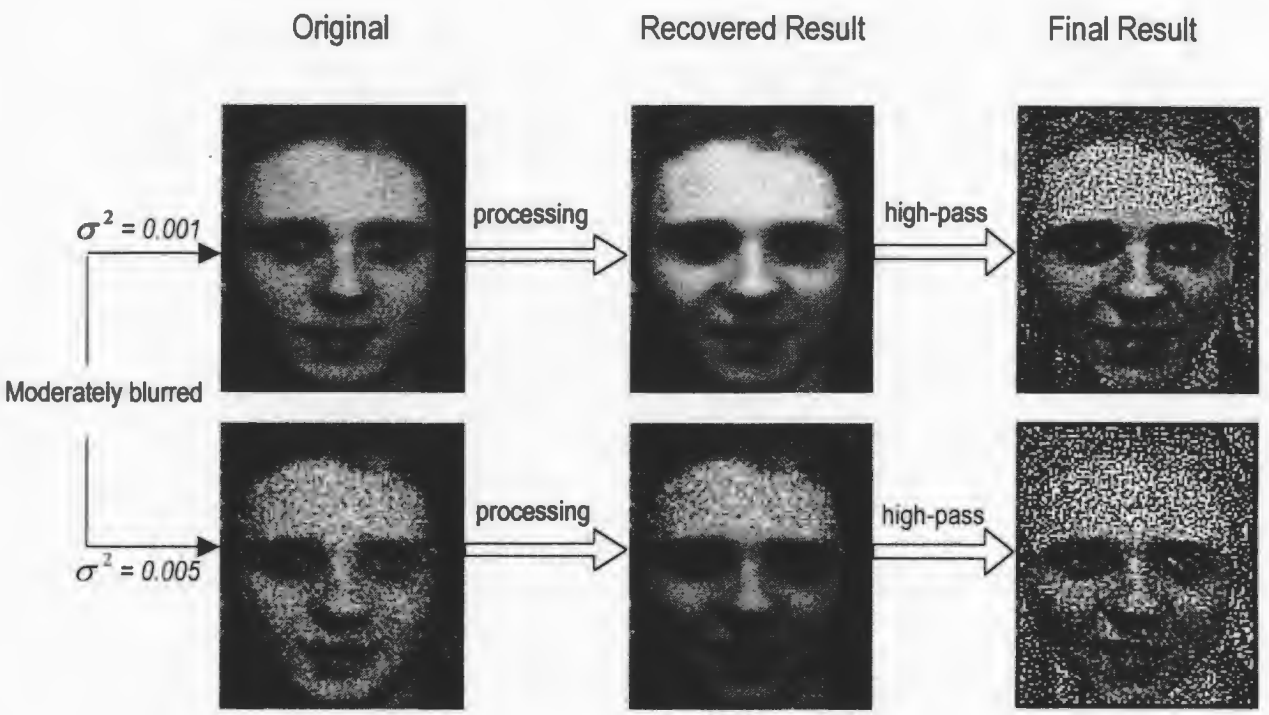
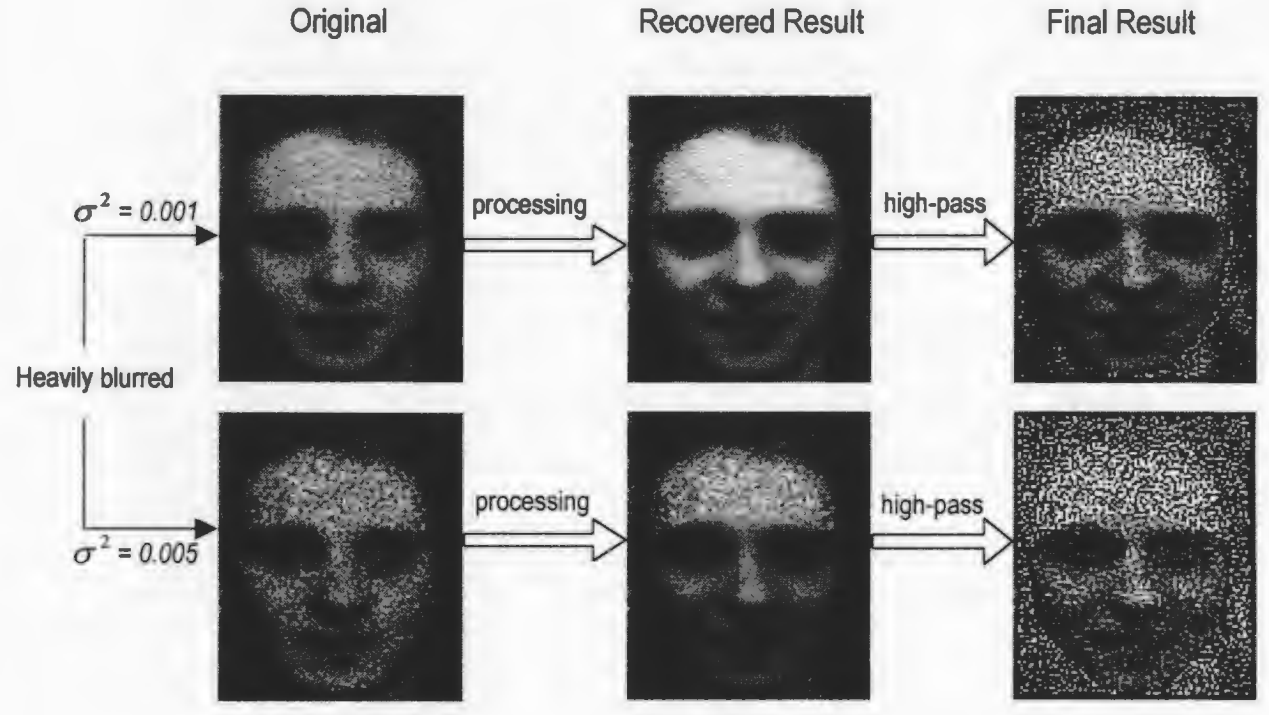


Fig.5.7 Recovered Target Frames from Noisy and Blurred Images (continued)





(b) the Second Candidate

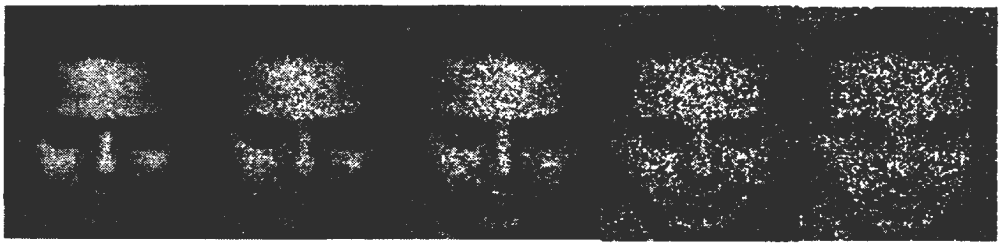
Fig.5.7 Recovered Target Frames from Noisy and Blurred Images (continued)

5.6 Comparison with Other Filters

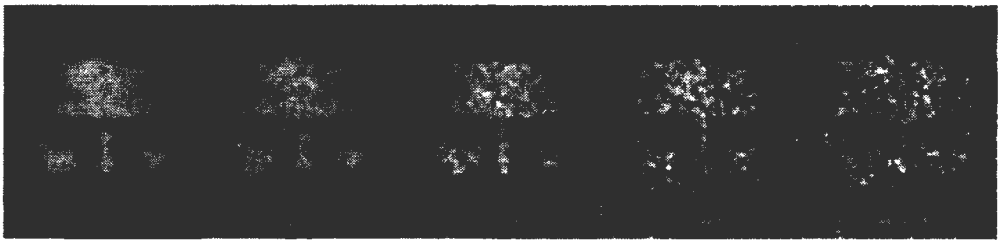
For comparison, the enhancement results from frames with different noise degrees, which are based on the proposed approach and the existing spatial filters such as median filter and MMSE filter, are listed in Fig.5.8.

The original noisy frames are displayed in Fig.5.8(a). The noise variance from left to right is *0.001*, *0.005*, *0.01*, *0.03* and *0.05* respectively. The corresponding enhancement results by the median filter, the MMSE filter and the proposed approach in this thesis are shown in Fig.5.8(b), Fig.5.8(c) and Fig.5.8(d) respectively. Here, the 2-D median filter with a mask size of 3 is used. For the MMSE filter, the mask size is 5.

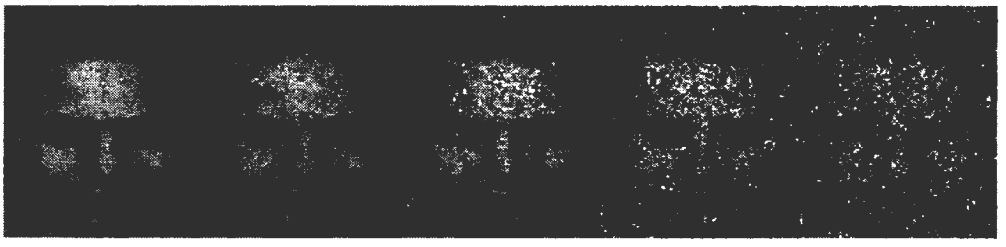
In Fig.5.8(b), the blurring effect caused by the median filter is disturbing. The MMSE filter works well when the noise variance is 0.001 and 0.005. However, if the noise variance is increased to 0.05, a blocking effect appears and facial details are missing. Compared with Fig.5.8(b) and Fig.5.8(c), the new approach in this thesis achieves best result for various noise variances.



(a) Original Noisy Target Frame



(b) Recovered Target Frame by Median Filter



(c) Recovered Target Frame by MMSE Filter

(d) Recovered Target Frame by the Proposed Approach in the Thesis

Fig.5.8 Comparison with Other Filters

Chapter 6 Subjective Tests

As described before, the foundation for all operations in the system is the locations of facial landmarks that are identified by a human being. Obviously, different sets of landmarks on the same frame will be generated by different users.

A supervised system will be impractical if the variation between users causes significantly different results. Therefore, it is necessary to test whether the system is sensitive to the landmark variations created by different users of the software.

In subjective tests, landmarks of 4 selected sequences from basic experiments are obtained by 9 inexperienced users. The similar results on each frame sequence obtained from different users as in Chapter 5 demonstrate the effectiveness and flexibility of the system.

6.1 Application for Permission

Because the participants in the experiment are human subjects, these experiments require the approval of *the Ethics Committee of the Faculty of Science*. The corresponding application documents are attached in *Appendix 3*.

6.2 Selecting Frame Sequences

It is expected that people's judgement on landmark locations will deviate more as the frame quality becomes worse. For example, people will have different judgements on the location of an eye corner given a heavily noisy frame while showing agreement on the same landmark if the frame looks clear. The intent of this experiment is to determine the

variation of landmark locations and recovery results by different operators. Therefore, the hard to detect sequences will be of most interest.

In the basic experiments in Chapter 5, the algorithm was tested on 24 frame sequences with varying degree of noise and blur for 2 candidates (see Section 5.2). For each candidate, there were 12 kinds of combinations of noise and blur summarized in Table 6.1.

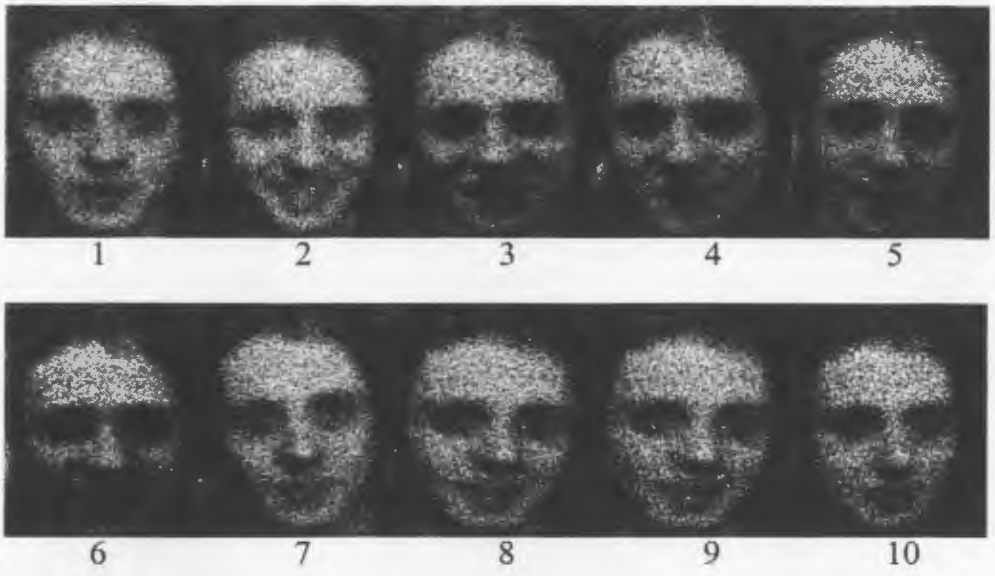
To save work, the easily distinguished frames were not included in the subjective tests. The 4 sequences in the shaded area in Table 6.1 are of relatively poor quality. These sequences with the degradation combination indexed 9, 11 for the second candidate and 6, 8 for the first candidate are selected in subjective tests (see Table 6.2 for details). All images are shown in Fig.6.1.

Index	Noise Variance	Blur Degree	Recovery Result
1	0	0	Fig.5.5
2	0.001	0	Fig.5.6
3	0.001	Slight	Fig.5.7
4	0.001	Moderate	Fig.5.7
5	0.001	Heavy	Fig.5.7
6	0.005	0	Fig.5.6
7	0.005	Slight	Fig.5.7
8	0.005	Moderate	Fig.5.7
9	0.005	Heavy	Fig.5.7
10	0.01	0	Fig.5.6
11	0.03	0	Fig.5.6
12	0.05	0	Fig.5.6

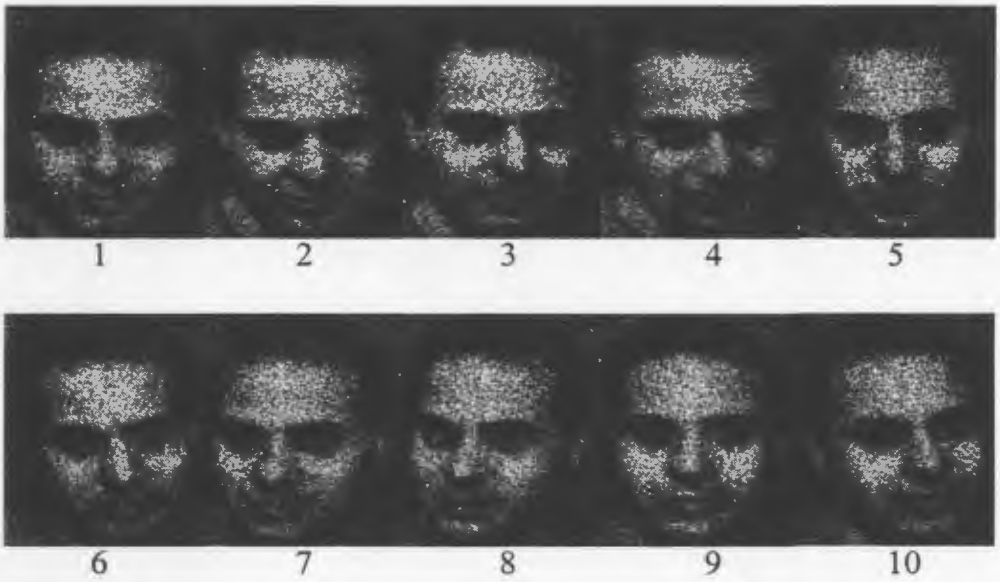
Table 6.1 Summary of Previous Experiments

Sequence	Description		
	candidate	noise variation	Blur degree
1	second	0.005	heavy
2	first	0.005	moderate
3	second	0.03	no blur
4	first	0.005	no blur

Table 6.2 Four Sequences for Subjective Tests

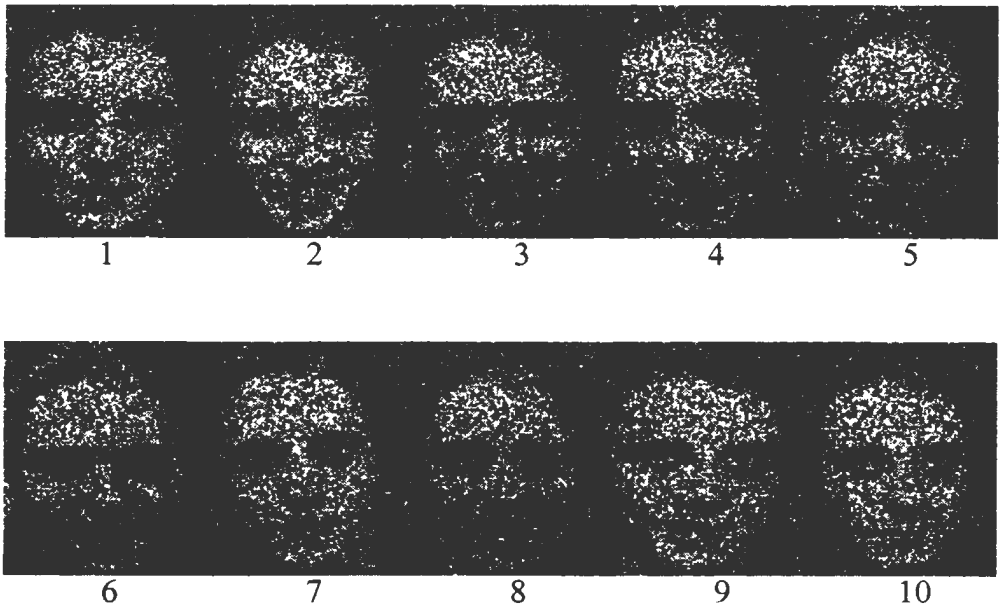


(a) Degraded Frames in Sequence 1

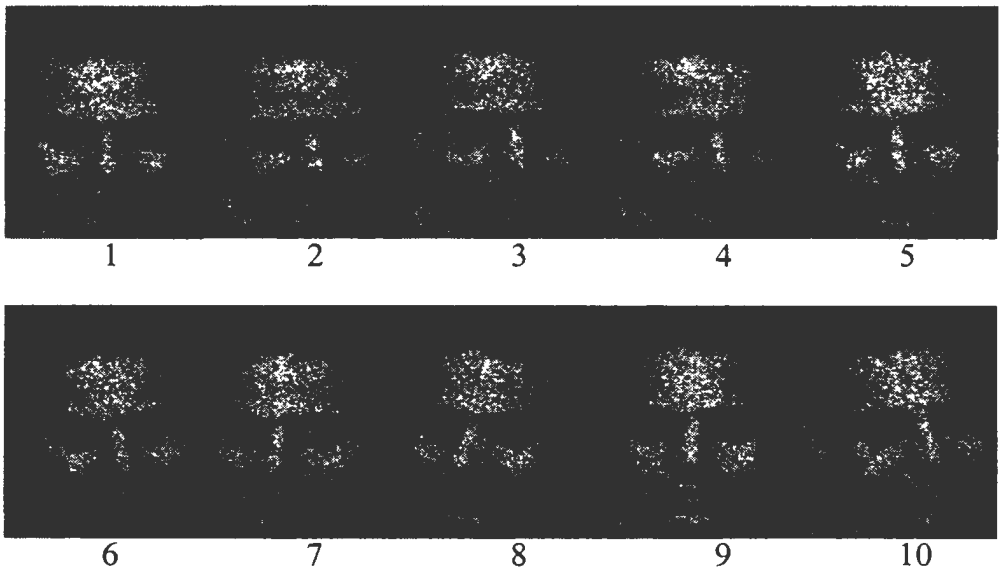


(b) Degraded Frames in Sequence 2

Fig.6.1 Four Degraded Sequences Selected for Subjective Tests



(c) Degraded Frames for Sequence 3



(d) Degraded Frames for Sequence 4

Fig.6.1 Four Degraded Sequences Selected for Subjective Tests (continued)

6.3 Experiment Description

Nine subjects who had no previous knowledge of the testing sequences participated in the experiment. Given 4 groups of frame sequence (up to 10 frames in each sequence), the subjects were required to identify the landmarks on each frame manually and save the locations of the landmarks in a data file in the corresponding directory.

Locating on the frame sequence in the worst case to be identified was performed first to avoid a priori knowledge on the same candidate. Each subject encountered the sequences in the same order as indicated in Table 6.2.

Because locating landmarks is the only operation performed by the subjects in the experiment, a simplified version of the interface introduced in Chapter 4 named the *Facial Landmark Locating Interface* is provided and shown in Fig.6.2. The functionality of the buttons is the same as the descriptions in Chapter 4.

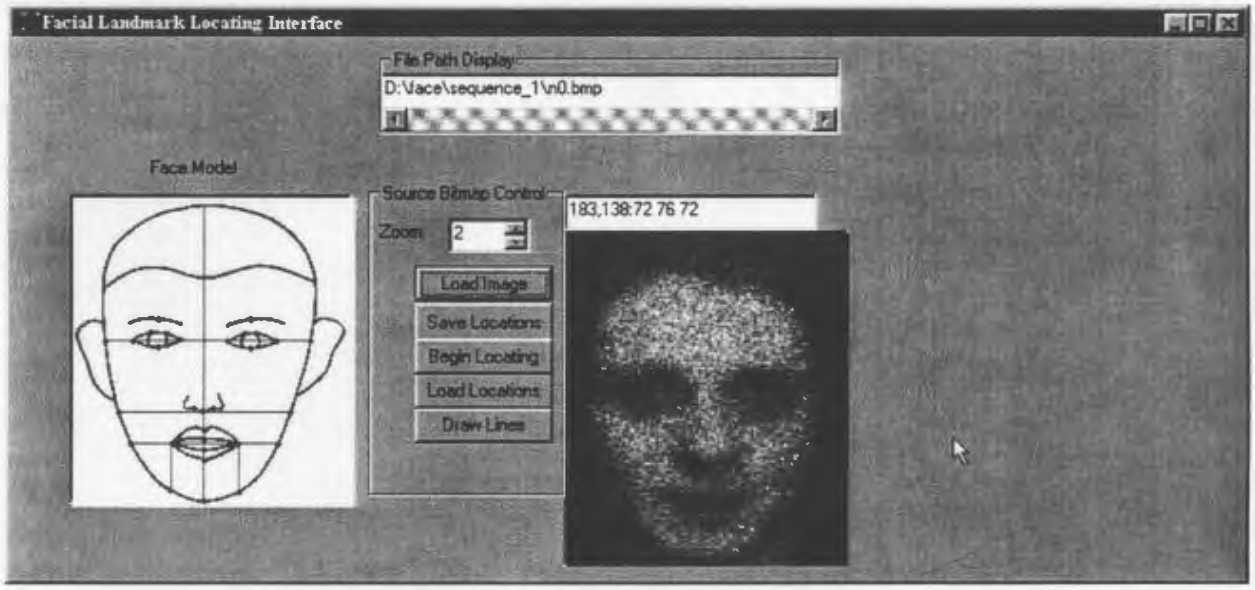


Fig.6.2 Facial Landmark Locating Interface for Subjective Tests

The work was done under this interface, which is an easy-to-use executable program running in Windows 95. All loaded frames were double-sized automatically. The working environment, including assigning the individual directory and instructing on the user interface was prepared by the author. The instruction manual presented to the subjects during the work is attached in *Appendix 4*.

The data files generated by each subject were used to recover de-noised facial frames. By comparing the facial frame recovered by each subject, it can be concluded whether our supervised system is feasible for enhancement of facial image sequences independent of the operator.

6.4 Analysis on Subjects' Recovery Results

Fig.6.3 shows the recovered target frames by 9 subjects. Fig.6.4 shows the ideal recovery result from degraded frames (see Section 6.4.1 for detailed definition on ideal recovery result). The frames numbered from 1 to 4 are recovered from sequence 1 to 4 respectively. Frames 5 and 6 are obtained by using de-blur filtering on the recovery results numbered 1 and 2 respectively.

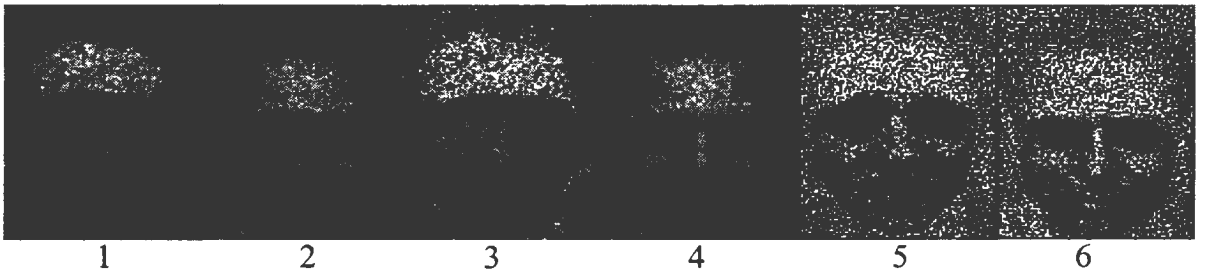
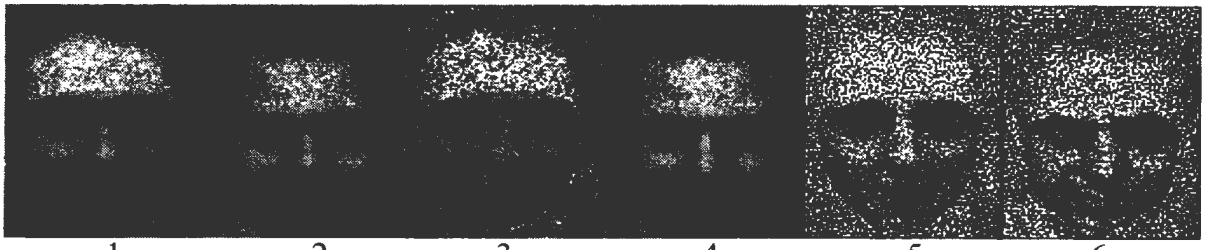
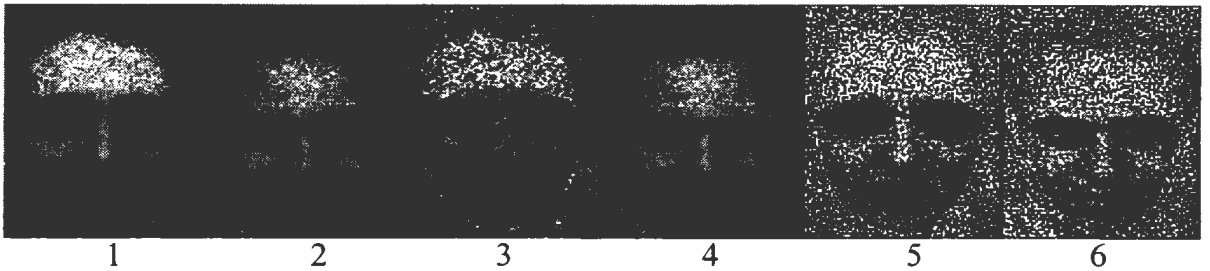


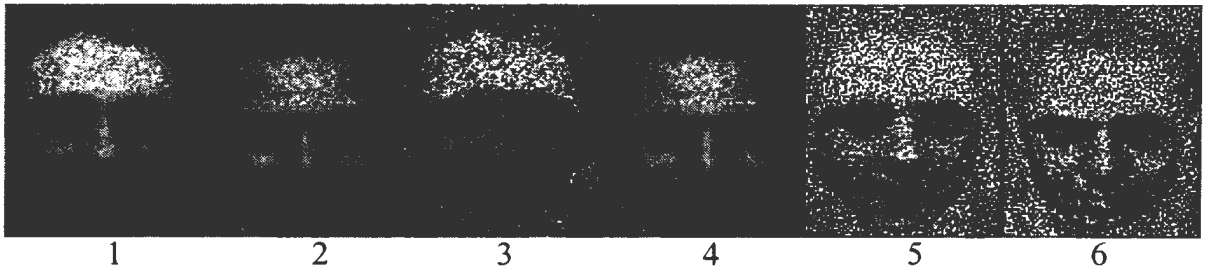
Fig. 6.3 Recovery Results by Subjects



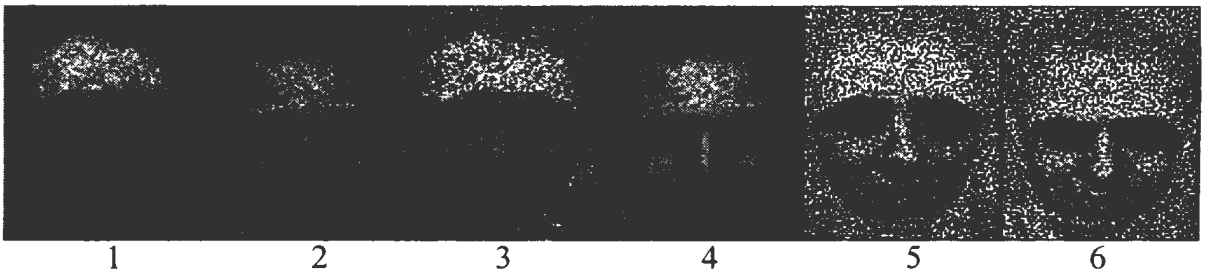
Recovery Results by Subject 2



Recovery Results by Subject 3

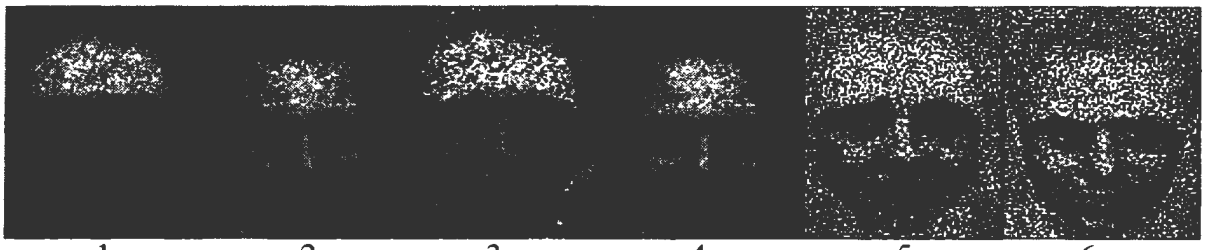


Recovery Results by Subject 4

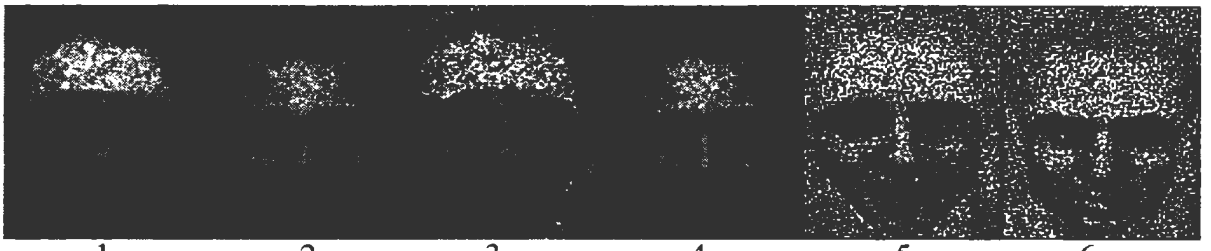


Recovery Results by Subject 5

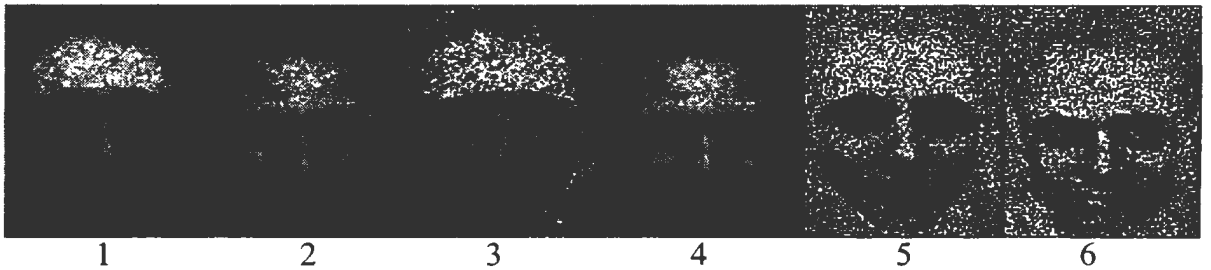
Fig. 6.3 Recovery Results by Subjects (continued)



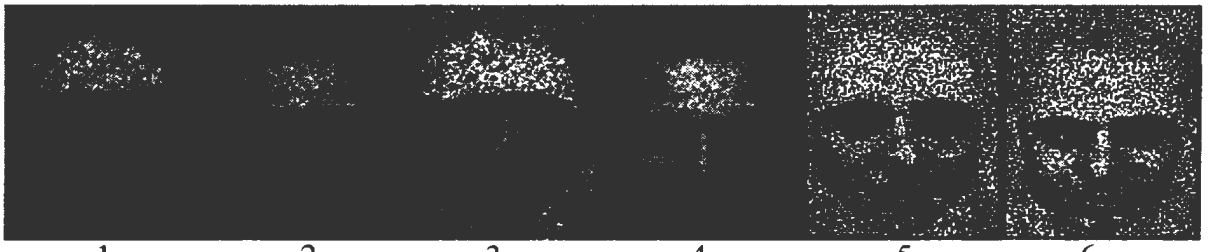
Recovery Results by Subject 6



Recovery Results by Subject 7



Recovery Results by Subject 8



Recovery Results by Subject 9

Fig.6.3 Recovery Results by Subjects (continued)

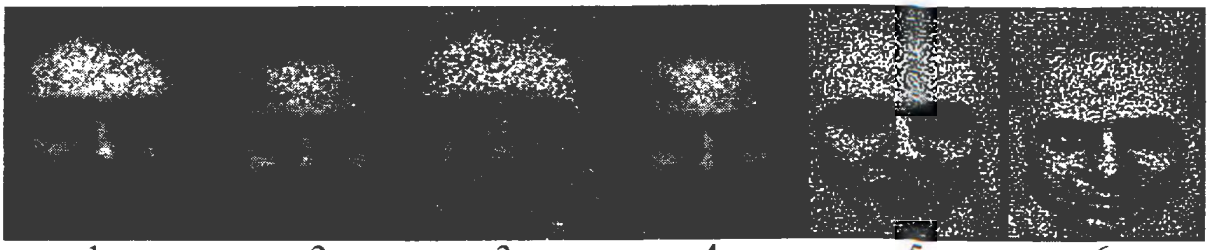


Fig.6.4 Ideal Recovery Results

Observing the recovery results in Fig.6.3 and Fig.6.4, the appearance of the ideal recovery results and the recovery results by different operators are similar. But the measurement should not made only by eye. Quantitative analysis is also required. Normally, it is done by MSE or PSNR calculation (see below for detailed definition). Although neither MSE nor PSNR necessarily correlates with subjective error, they are widely used in the absence of a widely-agreed subjective measure. In this thesis, the quantitative evaluation is achieved by PSNR calculation.

The PSNR (*Peak Signal to Noise Ratio*) value can fairly describe the grey level difference between two images of the same size. The formula for the calculation is listed in (6.1):

$$PSNR(dB) = 10\log_{10}\frac{X_{\max}^2}{MSE} \tag{6.1}$$

where $X_{\max}$ is the maxium possible grey level. In this experiment, $X_{\max}$ is equal to 255. MSE is *mean square error* defined in (6.2):

$$MSE = \frac{1}{N \times M} \sum_{\substack{i=1,N \\ j=1,M}} (X_{ij} - \hat{X}_{ij})^2 \tag{6.2}$$

where N is the number of columns and M is the number of rows. X_{ij} and $\hat{X}_{ij}$ are grey levels of a pixel at the same position $[i, j]$ in two images.

If the two images are exactly the same, the PSNR value will be infinity. If one image is completely dark with grey level 0 on all pixels and the other is completely white with grey level 255 on all pixels, we will have the minimum PSNR value: zero.

There is one special note about our PSNR calculation. The landmarks numbered 22, 24, 26, 28, 29, 30, 27, 25, 23, 32, 33 and 31 (see Fig.3.2) compose a polygon where frame averaging is performed. The quality of this processed face area is our concern and only grey level information of the pixels within this area of the two frames should contribute to PSNR. This area is called *the facial mask.*

Therefore, three basic components are desired when calculating PSNR: source image, destination image and the involved area. Source image and destination image are the two images for comparison. Facial mask selection determines the area involved in calculation. The conventions to define their instances are as follows:

1. Source image includes 5 instances: *subject's recovery result, subject's deblurred recovery result, ideal recovery result, deblurred ideal recovery result* and *original degraded target frame.* All recovery results of source image instances are from degraded frames (see Fig.6.1).

- *Subject's recovery result* refers to the frame averaging result of degraded frames based on landmark locations obtained from degraded frames by subjects.
- *Subject's deblurred recovery result* refers to the final recovery result obtained after processing with a high-pass filter on subject's recovery result from degraded frames.
- *Ideal recovery result* refers to the frame averaging result of degraded frames based on landmark locations obtained from clear frames. This can be assumed to be the best result based on the proposed approach.
- *Deblurred ideal recovery result* refers to the high-passed ideal recovery result.
- *Original degraded target frame* refers to the frames presented to subjects. The frames numbered 1 in Fig.6.1 are samples.

2. Destination image has 4 instances: *clear frame, noise-free frame, subject's recovery result from noise-free frames*, and *ideal recovery result from noise-free frames.*
   - *Clear frame* refers to the undegraded frame which is both noise-free and blur-free(see Fig.6.5 (c) and (d)).
   - *Noise-free frame* refers to the blur degradation result on original clear frames. For sequence 1, it refers to the heavily blurred noise-free frames as in Fig.6.5(a). For sequence 2, it refers to the moderately blurred noise-free frames as in Fig.6.5(b). For sequence 3 and 4, it has the same meaning as *clear frame.*
   - *Subject's recovery result from noise-free frames* refers to the frame averaging result of noise-free frames based on landmark locations obtained from degraded frames by subjects. It has the same shape as the subject's recovery result from degraded frames.

- *Ideal recovery result from noise-free frames* refers to the frame averaging result of noise-free frames based on landmark locations obtained from clear frames by subjects. It has the same shape as the ideal recovery result from degraded frames.

3. The shape of the recovered target frame depends on the mask of target frame which all other frames will be transformed to. Thus the facial mask for PSNR calculation is defaulted as the one for target frame. There are two kinds of mask selections:
   - *Clear mask* refers to the facial area which is bounded by landmarks obtained from the clear target frame. It is supposed to fit the real target face exactly and be objective.
   - *Subject's mask* refers to the facial area which is bounded by landmarks located by subjects on the degraded target frame. Its accuracy depends on subject's judgement and image quality. It is an estimation.

To have a comprehensive view of the recovered frames, PSNR results based on different groups of PSNR components are calculated. The PSNR values and definitions of basic PSNR component combinations are reported in Table 6.3 to Table 6.4.

In Table 6.3, the variation of noise levels contained in recovery results is given. The source frame and the destination frame have the same shape because they are obtained by processing the degraded sequence and the noise-free sequence in the same way from subject's mask. Hence the only difference between the two frames is whether noise exists or not. This is also the measurement employed in Table 5.1. It can be read from Table 6.3 that the PSNR of all recovery frames have been increased by about 12 dB, which depends on the number of frames involved in frame averaging. In the experiment,

10 frames are averaged, which suggests a 10 dB improvement from discussion in section 2.1.3.

Noise information in the recovered target frame can be measured accurately in this way, but the subject's recovery result from noise-free frames is not objective for quality comparison. Obviously, the clear target frame provides objective comparison. In Table 6.4, the clear target frame is used to compare with the subjects' deblurred recovery results for sequence 1 and 2 and recovery results for sequence 3 and 4 on clear mask. The average improvement of PSNR for sequence 1, 2, 3 and 4 is 6.3376, 6.4288, 4.4664 and 3.7025.

The PSNR value for the ideal results is also available in Table 6.3 and 6.4. In Table 6.3, the ideal recovery result from degraded frames is compared with the one from the noise-free frame and the clear target frame. The deblurred ideal recovery result from degraded frames is compared with the one from noise-free and clear target frame. It is deducted that no significant variation between subjects' recovery result and the ideal result exists based on PSNR value.

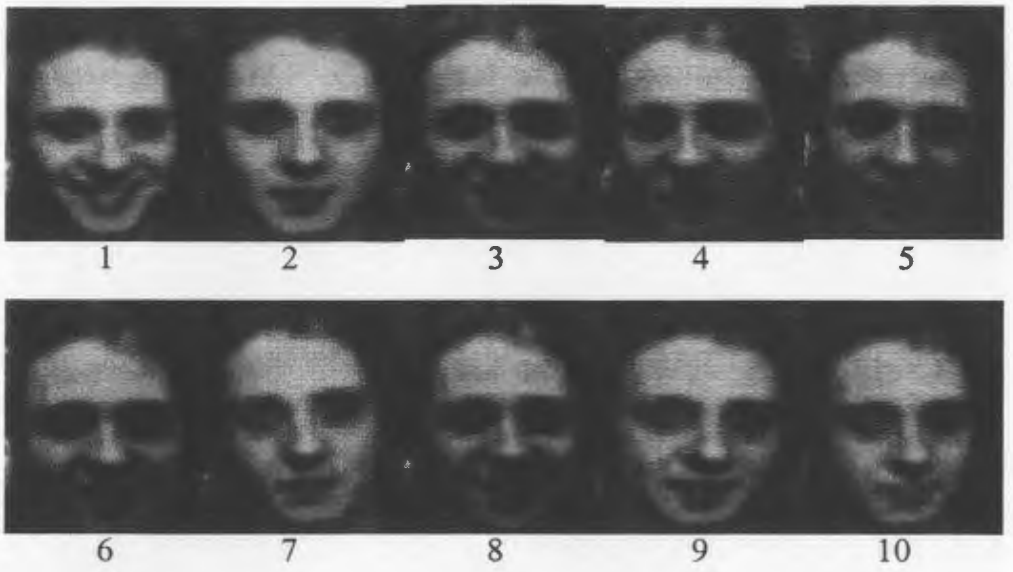
The clear mask is used in Table 6.4. Because the two frames for comparison do not have the same shape, the face area of the destination frame and the source frame do not match. Therefore, PSNR improvement is less than presented in Table 6.3 where both frames have exactly the same shape as the subjects' mask. The slight displacement between two frames will cause much degradation in PSNR, especially for frames full of details.

From pictures in Fig.6.3 and Fig.6.4, the luminance of the recovery results are different from the original degraded target frames (see Fig.6.1). This is because all frames in the sequence which are in different lighting conditions contribute to the recovery result.

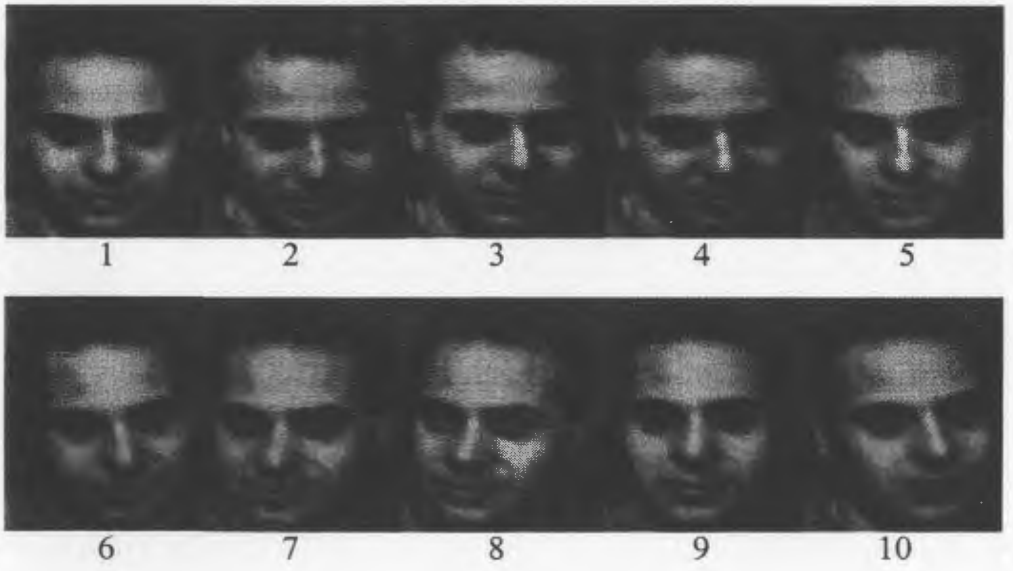
Histogram equalization (see Section 2.1.1) is an effective method to obtain uniform luminance. Fig.6.6 and Fig.6.9 show the results of histogram equalization on the original degraded frames and noise-free/clear frames respectively. If frame averaging is performed on frames in Fig.6.6, the luminance difference between recovered results (see Fig.6.7 and Fig.6.8) and noise-free/clear frames is reduced. The previous experiments (see Table 6.3 and Table 6.4) are repeated again based on equalized frames. The corresponding results are reported in Table 6.5 and Table 6.6.

When the effect of luminance is removed, better quality recovered frames are expected because of the contrast enhancement. However, noise is also amplified. This brings about the degradation for most PSNR measurements.

In addition, the insignificant variance among subjects presented from Table 6.3 to Table 6.6 verifies the similarity of the recovery results obtained by different operators.

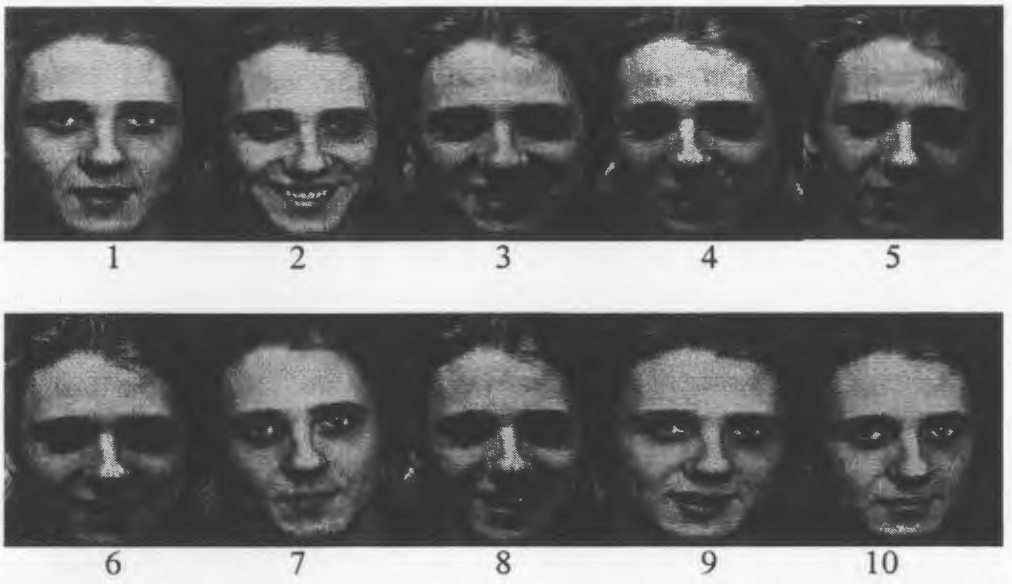


(a) Noise-free Frames of Sequence 1

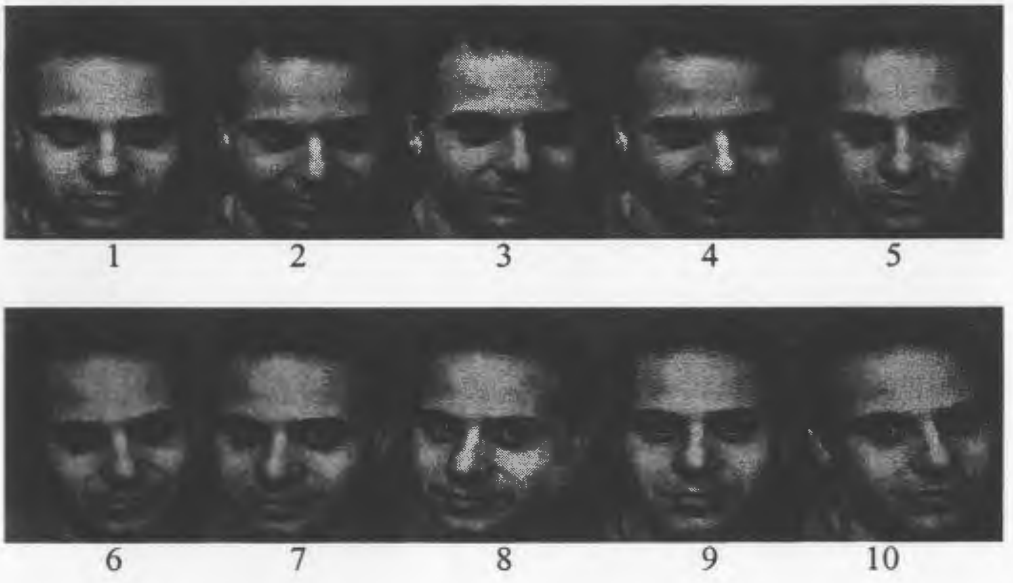


(b) Noise-free Frames for Sequence 2

Fig.6.5 Noise-free/Clear Frames in Subjective Tests



(a) Clear Frames for Sequence 1, Noise-free/Clear Frames for Sequence 3



(b) Clear Frames for Sequence 2, Noise-free/Clear Frames for Sequence 4

Fig.6.5 Noise-free/Clear Frames in Subjective Tests (continued)

Source	Sequence 1		Sequence 2		Sequence 3		Sequence 4	
	Original	Recov.	Original	Recov.	Original	Recov.	Original	Recov.
Ideal	23.0882	35.6066	22.9686	35.5364	14.0689	25.1316	22.9686	35.5634
Recov. by Subject								
1	23.0688	35.1955	23.0379	35.4135	14.2547	25.5853	23.0087	35.3403
2	23.0475	35.3982	23.0028	35.5404	14.0761	25.2310	23.0202	35.2916
3	23.0479	35.5302	22.9987	35.3482	14.0061	24.9823	23.0166	35.4052
4	23.0742	35.4307	22.9747	35.2989	14.1390	25.3468	23.0119	35.3243
5	23.0693	35.5041	23.0117	35.2518	14.2196	25.3889	22.9850	35.5314
6	23.0603	35.3578	22.9996	35.3515	14.0760	25.1287	22.9964	35.3535
7	23.0734	35.4317	23.0087	35.3126	14.0914	25.1447	23.0323	35.1205
8	23.0794	35.3434	22.9968	35.3610	14.2280	25.5458	23.0090	35.2871
9	23.0674	35.4215	23.0190	35.2723	14.1742	25.4470	23.0031	35.3165
Average	23.0654	35.4015	23.0055	35.3500	14.1406	25.3112	23.0092	35.3300
Variation	0.00011	0.00855	0.00027	0.00668	0.00632	0.03684	0.00017	0.01048
Improve-ment	12.3361		12.3445		11.1706		12.3208	

Recovery PSNR:

Source image: subject's recovery result or ideal recovery result from degraded frames.

Destination image: subject's recovery result or ideal recovery result from noise-free frames. For sequence 1 and 2, blur degradation exists in noise-free frames.

Involved area: subject's mask on target frame.

Original PSNR:

Source image: original degraded target frame.

Destination image: original noise-free target frame. Note that blur degradation exists for sequence 1 and 2.

Involved area: subject's mask on target frame.

Average: the average PSNR by 9 subjects.

Improvement: the average PSNR improvement of 9 subjects.

Variation: the variation of PSNR among 9 subjects.

Table 6.3 PSNR Report 1

Source	Sequence 1	Sequence 2	Sequence 3	Sequence 4
Original	8.0341	8.18797	14.3828	22.9686
Ideal Recovery	15.4552	15.9619	19.8238	27.1872
Recovery by Subject				
1	14.1229	14.5052	18.4113	26.2466
2	13.9469	14.9563	19.0985	27.3971
3	14.4728	13.7949	18.9096	27.2878
4	14.3523	14.4772	18.9464	26.7401
5	14.5109	14.2598	18.3717	26.7679
6	14.4008	14.7131	19.3091	26.1307
7	14.2989	14.9408	19.4339	26.3020
8	14.5682	14.7031	18.6309	26.7544
9	14.6718	15.2009	18.531	26.4131
Average	14.37172	14.61681	18.84916	26.67108
Improvement	6.33762	6.42884	4.46636	3.70248
Variation	0.045239	0.157005	0.133033	0.178166

Recovery PSNR:

Source image: For sequence 1 and 2, subject's deblurred recovery result or deblurred ideal recovery result.
For sequence 3 and 4, subject's recovery result from degraded frames or ideal recovery result.

Destination image: original clear target frame. Note that blur degradation does not exist for sequence 1 and 2.

Involved area: clear mask on target frame.

Original PSNR:

Source image: For sequence 1and 2, original deblurred degraded target frame.
For sequence 3 and 4, original degraded target frame.

Destination image: original clear target frame. Note that blur degradation does not exist for sequence 1 and 2.

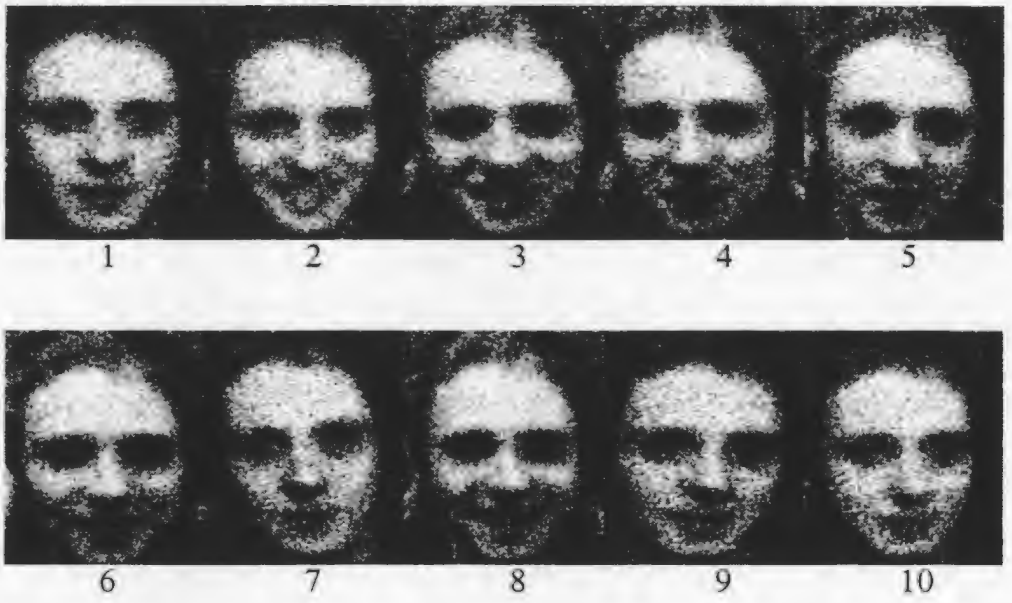
Involved area: clear mask on target frame.

Average: the average PSNR by 9 subjects.

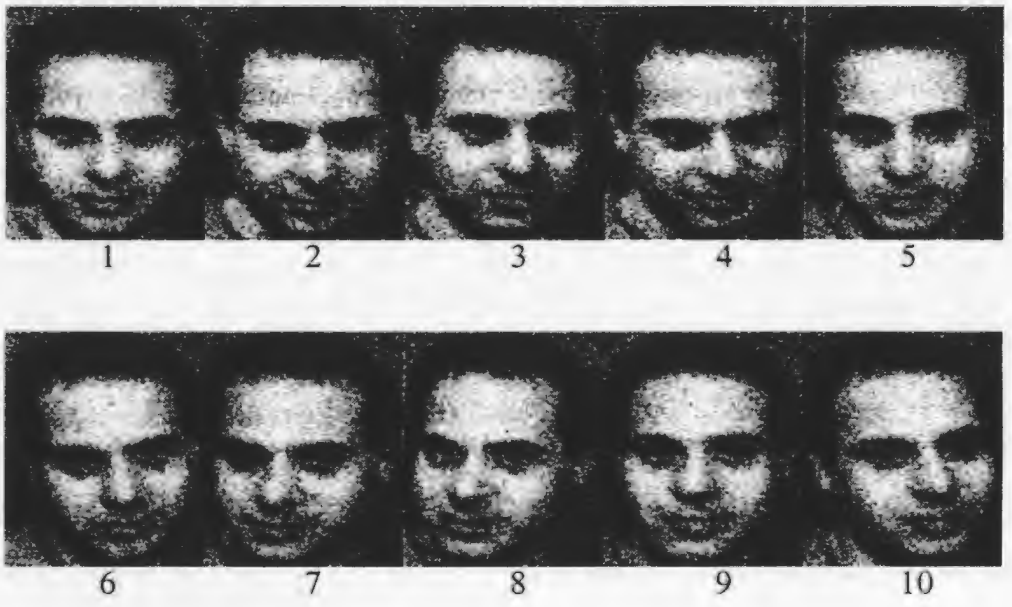
Improvement: the average PSNR improvement of 9 subjects.

Variation: the variation of PSNR among 9 subjects.

Table 6.4 PSNR Report 2

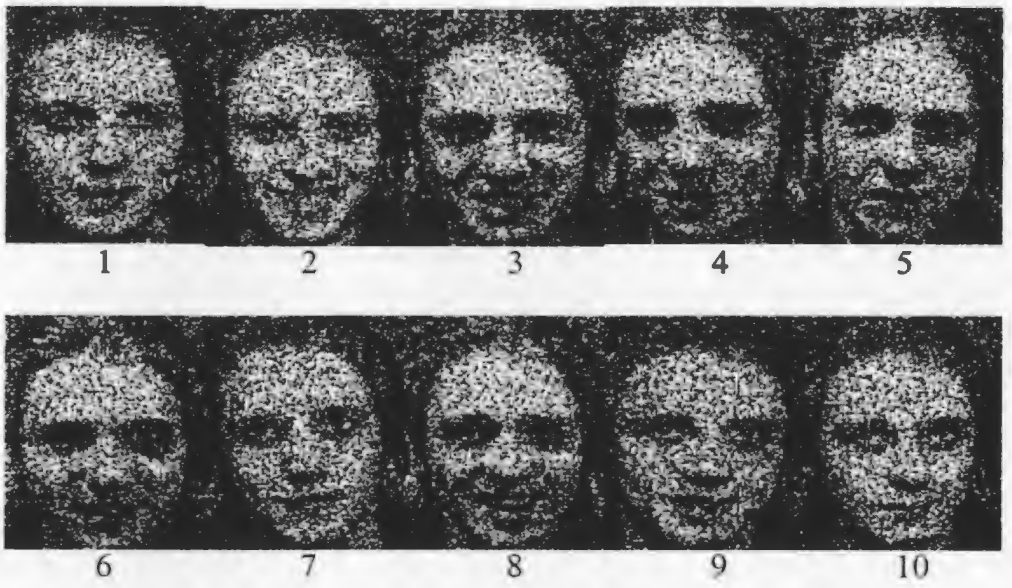


(a) Degraded Frames for Sequence 1

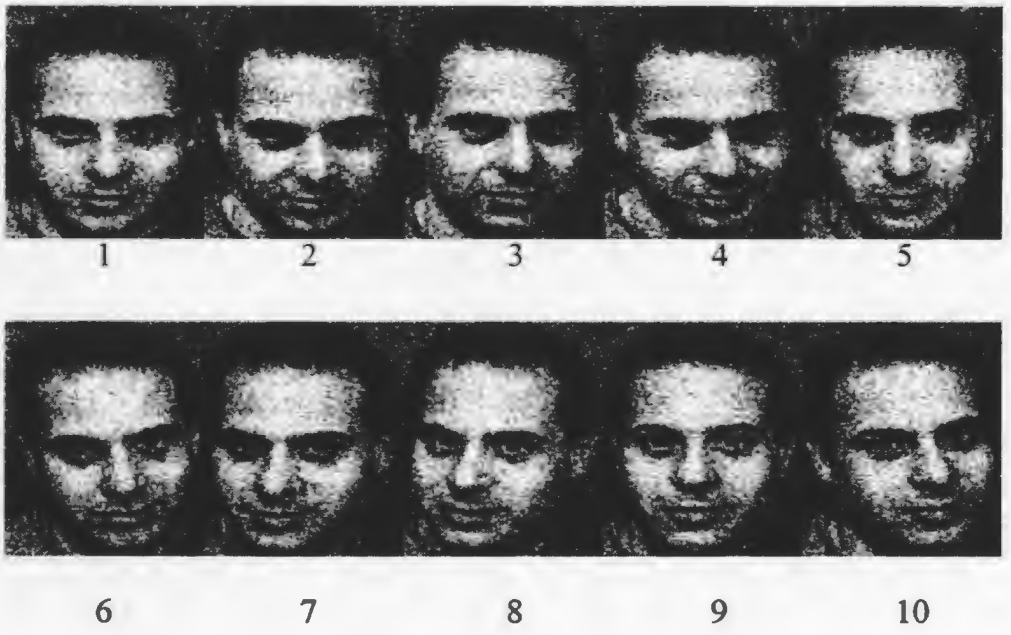


(b) Degraded Frames for Sequence 2

Fig.6.6 Four Degraded Sequences Selected for Subjective Tests
(After Histogram Equalization)



(c) Degraded Frames for Sequence 3



(d) Degraded Frames for Sequence 4

Fig.6.6 Four Degraded Sequences Selected for Subjective Tests (Continued)
(After Histogram Equalization)

Recovery Results by Subject 1

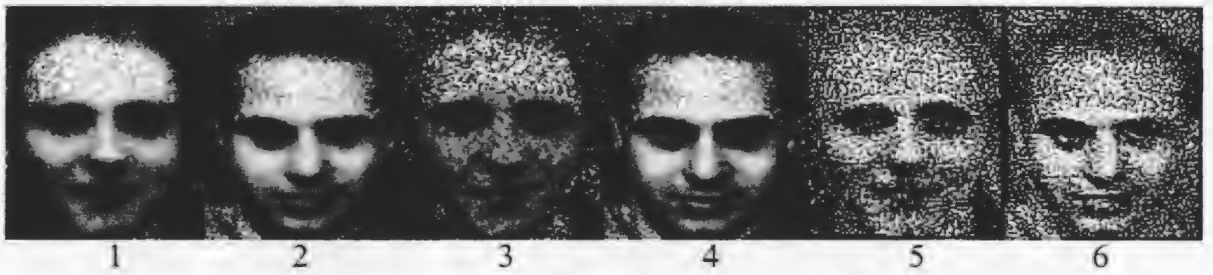
Recovery Results by Subject 2

Recovery Results by Subject 3

Recovery Results by Subject 4

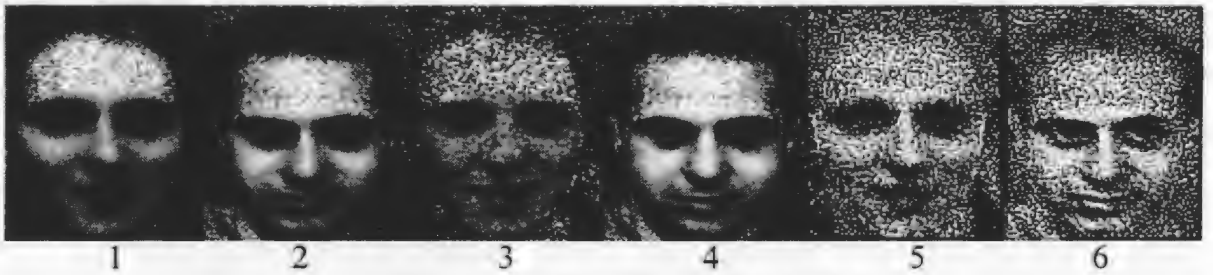
Fig.6.7 Recovery Results by Subjects
(After Histogram Equalization)

Recovery Results by Subject 5



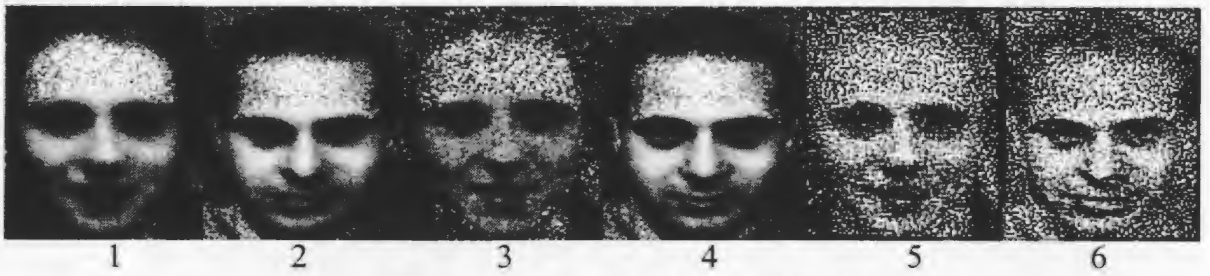
Recovery Results by Subject 6

Recovery Results by Subject 7



Recovery Results by Subject 8

Fig.6.7 Recovery Results by Subjects (Continued)
(After Histogram Equalization)



Recovery Results by Subject 9

Fig.6.7 Recovery Results by Subjects (Continued)
(After Histogram Equalization)

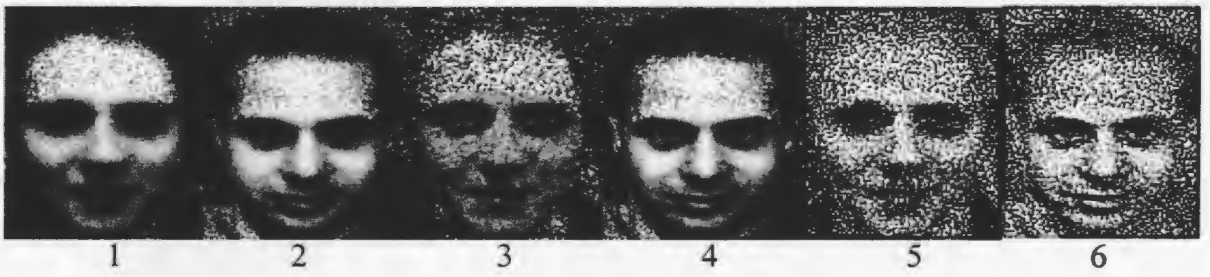
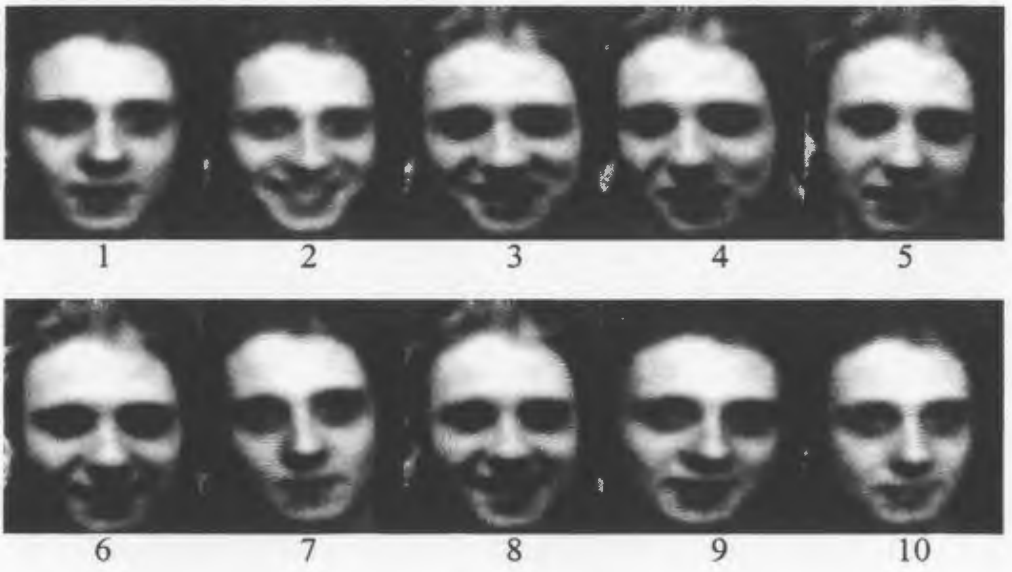
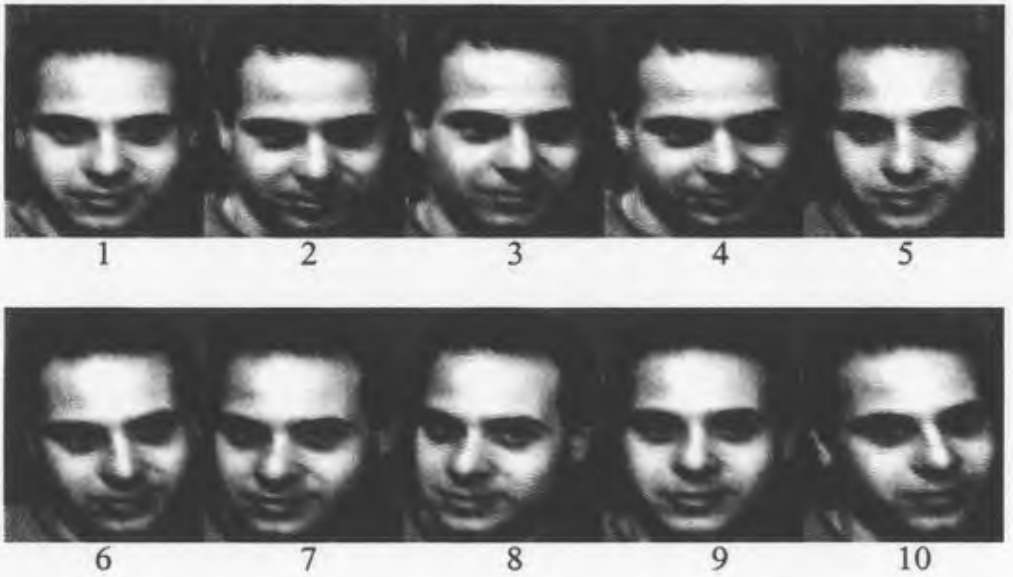


Fig.6.8 Ideal Recovery Results (After Histogram Equalization)

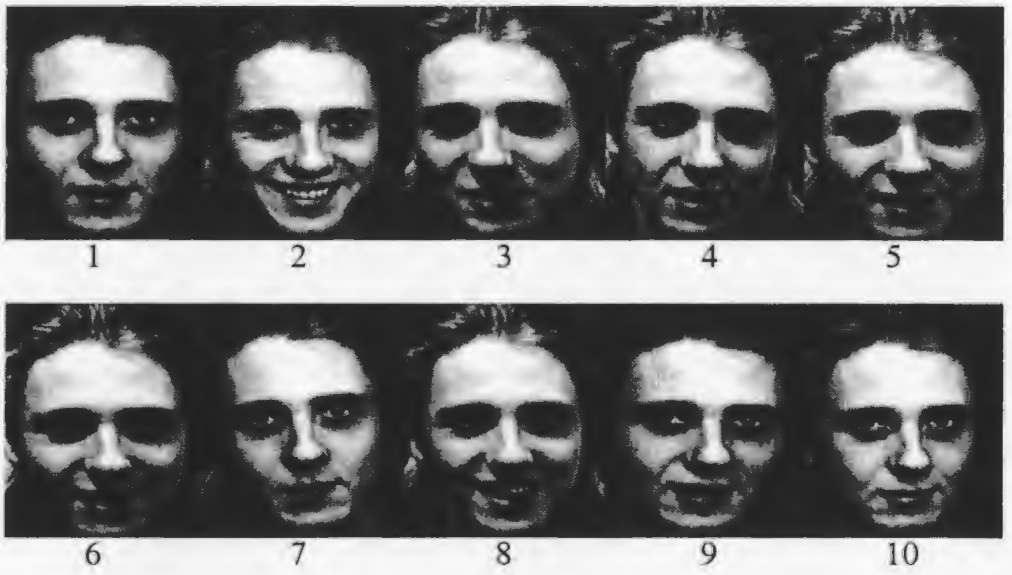


(a) Noise-free Frames of Sequence 1

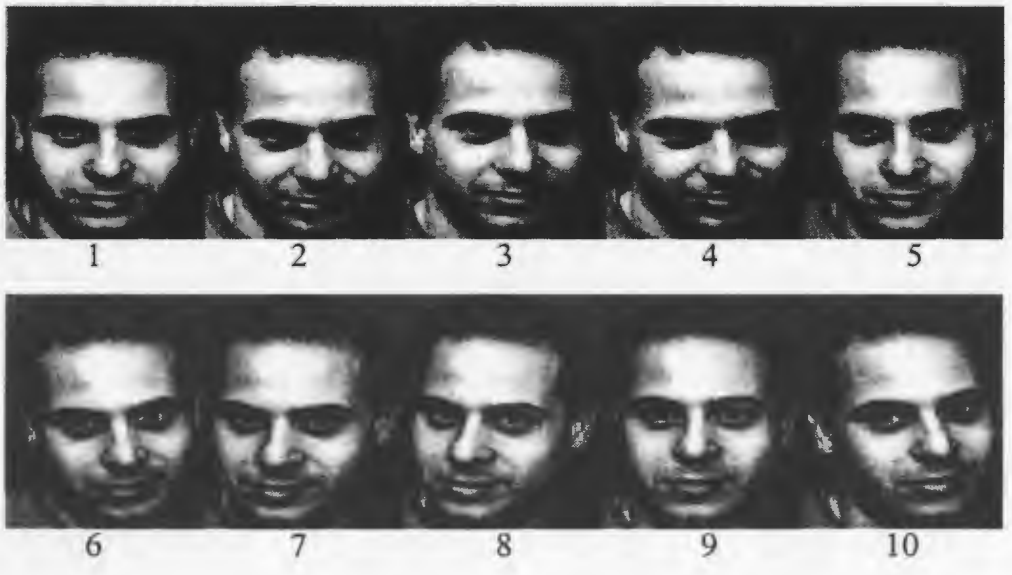


(b) Noise-free Frames of Sequence 2

Fig.6.9 Noise-free/Clear Frames in Subjective Tests
(After Histogram Equalization)



(c) Clear Frames for Sequence 1
Noise-free/Clear Frames for Sequence 3



(d) Clear Frames for Sequence 2
Noise-free/Clear Frames for Sequence 4

Fig.6.9 Noise-free/Clear Frames in Subjective Tests (Continued)
(After Histogram Equalization)

Source Data	Sequence 1		Sequence 2		Sequence 3		Sequence 4	
	Original	Recov.	Original	Recov.	Original	Recov.	Original	Recov.
Ideal	19.2766	30.4250	19.4976	30.4530	12.6085	21.0297	19.8754	30.9121
Subject								
1	19.2276	30.3488	19.4411	30.7704	12.7902	21.7934	19.8400	30.9170
2	19.2337	30.5210	19.5692	30.6891	12.6278	21.3516	19.8634	30.8711
3	19.2367	30.7849	19.6081	30.6205	12.5761	21.1390	19.8346	30.9878
4	19.2544	30.6952	19.4771	30.5319	12.6737	21.4232	19.8613	30.9639
5	19.2563	30.6825	19.5105	30.4709	12.7676	21.5535	19.8661	31.0788
6	19.2215	30.5553	19.4859	30.5097	12.607	21.3046	19.8186	30.9603
7	19.2497	30.5660	19.4984	30.5129	12.6236	21.2889	19.8701	30.7801
8	19.2502	30.6104	19.4750	30.5226	12.7711	21.7242	19.8444	30.9365
9	19.2357	30.4373	19.4524	30.4085	12.7026	21.6596	19.8370	30.9365
Average	19.24064	30.57793	19.50197	30.55961	12.68219	21.47089	19.84839	30.93689
Variation	0.000137	0.016083	0.002626	0.01138	0.005628	0.044227	0.000275	0.005924
Improve-ment	11.33729		11.05764		8.7887		11.0885	

Recovery PSNR:

Source image: subject's recovery result or ideal recovery result from degraded frames.

Destination image: subject's recovery result or ideal recovery result from noise-free frames. For sequence 1 and 2, blur degradation exists in noise-free frames.

Involved area: subject's mask on target frame.

Original PSNR:

Source image: original degraded target frame.

Destination image: original noise-free target frame. Note that blur degradation exists for sequence 1 and 2.

Involved area: subject's mask on target frame.

Average: the average PSNR by 9 subjects.

Improvement: the average PSNR improvement of 9 subjects.

Variation: the variation of PSNR among 9 subjects.

Table 6.5 PSNR Report 3 (After Histogram Equalization)

	Sequence 1	Sequence 2	Sequence 3	Sequence 4
Original	7.02906	7.40423	12.6085	19.8754
Ideal Recovery	12.7832	13.4576	18.7370	23.2084
Recovery by Subject				
1	12.1549	12.6726	17.2680	22.3471
2	11.9129	12.8925	17.8413	23.4365
3	12.1864	12.1867	17.6740	23.3355
4	12.0705	12.5206	17.8821	22.8465
5	12.4166	12.4637	17.1424	22.9883
6	12.1685	12.7687	18.0631	22.2952
7	12.3018	12.9461	18.2470	22.4314
8	12.4465	12.6167	17.4469	22.8744
9	12.4876	13.1569	17.2956	22.5810
Average	12.23841	12.69161	17.65116	22.79288
Improvement	5.209351	5.287381	5.042656	2.917478
Variation	0.032209	0.074406	0.132058	0.153151

Recovery PSNR:

Source image: For sequence 1 and 2, subject's deblurred recovery result or deblurred ideal recovery result.
For sequence 3 and 4, subject's recovery result from degraded frames or ideal recovery result.

Destination image: original clear target frame. Note that blur degradation does not exist for sequence 1 and 2.

Involved area: clear mask on target frame.

Original PSNR:

Source image: For sequence 1and 2, original deblurred degraded target frame.
For sequence 3 and 4, original degraded target frame.

Destination image: original clear target frame. Note that blur degradation does not exist for sequence 1 and 2.

Involved area: clear mask on target frame.

Average: the average PSNR by 9 subjects.

Improvement: the average PSNR improvement of 9 subjects.

Variation: the variation of PSNR among 9 subjects.

Table 6.6 PSNR Report 4 (After Histogram Equalization)

6.5 Variation of the landmark locations

In subjective tests, 4 sequences are selected. Each sequence includes 10 frames. Thirty landmarks need to be located manually on each frame. Thus there are altogether 1200 landmarks to be located by every subject. Subjects have different judgements on each landmark's position. The variation of landmark locations by subjects will be examined in this section.

Here, the mean value of x coordinate and y coordinate are represented by $\bar{x}$ and $\bar{y}$, the accurate landmark position obtained from clear frames are represented by $\hat{x}$ and $\hat{y}$, $\sigma_1^2(x)$ and $\sigma_1^2(y)$ are the variances from the mean value for x coordinate and y coordinate, $\sigma_2^2(x)$ and $\sigma_2^2(y)$ are the variances from the accurate value for x coordinate and y coordinate, and $\mathrm{cov}_1(x,y)$ and $\mathrm{cov}_2(x,y)$ are covariances from mean value and accurate value of x coordinate and y coordinate respectively. Specific definitions for $\bar{x}$ and $\bar{y}$, $\sigma_1^2(x)$, $\sigma_1^2(y)$, $\sigma_2^2(x)$ and $\sigma_2^2(y)$, $\mathrm{cov}_1(x,y)$ and $\mathrm{cov}_2(x,y)$ are shown in equations from (6.1) to (6.8), where N is the number of subjects who participated in the test (N is equal to 9 in this case) and $[x_i, y_i]$ is the coordinate for the same landmark by the *i*th subject.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6.1}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{6.2}$$

$$\sigma_1^2(x) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{6.3}$$

$$\sigma_1^2(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{6.4}$$

$$\sigma_2^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2 \tag{6.5}$$

$$\sigma_2^2(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{6.6}$$

$$\mathrm{cov}_1(x,y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\{(x_i - \bar{x})\times(y_j - \bar{y})\} \tag{6.7}$$

$$\mathrm{cov}_2(x,y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\{(x_i - \hat{x})\times(y_j - \hat{y})\} \tag{6.8}$$

In *Appendix 5.1* and *Appendix 5.2*, variance and covariance for landmarks in a particular testing frame are provided. Charts in *Appendix 5.1* are for $\sigma_1^2(x), \sigma_1^2(y)$ and $\mathrm{cov}_1(x,y)$. Charts in *Appendix 5.2* are for $\sigma_2^2(x)$, $\sigma_2^2(y)$ and $\mathrm{cov}_2(x,y)$. Frames indexed 1 to 10 are for sequence 1, indexed 11 to 20 for sequence 2, indexed 21 to 30 for sequence 3, and indexed 31 to 40 for sequence 4 (see Fig.6.1).

From charts in *Appendix 5.1* and *Appendix 5.2*, the following conclusions can be deducted:

- As well as image quality, deviation on landmarks is also affected by face expression and orientation. Variation for the mouth line (landmark 26 and 27) and nose line (landmark 24 and 25) is especially sensitive to orientation of the facial frame. In the 10th frame of sequence 2 (see Fig.6.1(b)) where vigorous out-of-plane rotation exists, the x-coordinate of landmark 26 (left side of nose

line) has the variance of 14.099 pixels. This is the maximum variance in *Appendix 5.1*.

- The variance in x-axis direction is always more than the variance in y-axis direction.
- The covariance of x coordinate and y coordinate is so small that x and y are almost decorrelated.

According to the definitions, it is reasonable to assume that the variances presented in *Appendix 5.1* indicate the confidence of operators during the locating process. The more confidence, the less variance among operators. Therefore, the landmarks with significant variance imply strong ambiguity during locating process. This brings about the future work to increase confidence for such landmarks by defining in a more accurate way or generating reference lines in the interface (refer to Chapter 8).

Meanwhile, it is also a reasonable assumption that the accuracy degree of landmark positions is reflected from the variances presented in *Appendix 5.2*. For example, landmark 14 (nose tip) has relatively small variances among overall frames in both *Appendix 5.1* (no more than 3 pixels) and *Appendix 5.2* (no more than 10 pixels). Such kind of landmarks with high confidence and accuracy (e.g. landmarks 14, 16, 18 and 29) can be used to help generate reference lines.

6.6 System Efficiency

Landmark positions obtained manually are the starting point of system. The efficiency of the algorithm depends heavily on the time spent on this manual work.

Apparently, the worse the image quality is, the more time will be spent on landmark locating. The average time spent for each frame in a selected sequence in subjective tests by 9 operators is listed in Table 6.7. The overall average time for the first sequence is 4.08 minutes per frame. For the second sequence, the overall average time is 3.01 minutes per frame. For the third and fourth sequence, the overall average time is 2.54 and 1.95 minutes per frame. It is therefore possible to process a 10-frame sequence for enhancement, achieving results similar to those shown earlier, in about 50 minutes.

Unit: minute/frame

Subject	Sequence No.			
	1	2	3	4
1	5.66	4.25	1.66	1.22
2	2.11	1.77	1.66	1.22
3	3.62	2.33	2.33	1.77
4	4.00	2.85	1.85	1.88
5	3.87	2.85	2.88	2.00
6	5.00	3.44	3.00	3.44
7	3.00	3.33	2.40	2.00
8	4.88	3.42	3.87	1.88
9	4.57	2.88	3.22	2.11
Overall Average	4.08	3.01	2.54	1.95

Table 6.7 Average Time Per Frame in Subjective Tests

Chapter 8 Conclusions, Contributions and Future Work

8.1 Conclusions

This thesis investigates the application of 2-D object modelling to image enhancement and restoration. A supervised system and its implementation were described in Chapter 3 and 4. To test the effectiveness of this strategy, some experiments were designed and performed in Chapter 5, Chapter 6 and Chapter 7.

Experiment 1: recovery from perfect sequences (Section 5.3)

In this experiment, the high quality images were used as input to the system. The results showed that no visible artifacts were introduced by the process.

Experiment 2: recovery from noisy and blurred sequences (Section 5.4 and Section 5.5)

The system was applied to simulated practical videos by degrading the high quality images from experiment 1. Different degrees of image degradation composed by noise and blur were examined. From 10 degraded frames, the system generated recovered images with PSNR improvements about 10 dB.

Experiment 3: recovery from practical video sequences (Chapter 7)

The frame sequences in this experiment were taken from videos of a simulated break-in. Subjective improvement was demonstrated.

Experiment 4: variation between users (Chapter 6)

Because of the importance of the landmark placement, whether the supervised system is sensitive to the landmark variations created by different users of the system was tested. The quantitative comparisons showed that there was only little difference between recoveries by different operators.

The results from all the above experiments demonstrated that supervised image warping and temporal filtering allows the enhancement and recovery of facial images from noisy and blurred videos. The introduced system is efficient, flexible, and practical.

- **Efficiency**

From the subjective tests in Chapter 6, the supervised process achieves substantial improvement within about 50 minutes of operator time for ten frames, and operator variation has minimal effect on recovered picture quality.

- **Flexibility**

The system accommodates a wide range of faces in different expressions, sizes, orientations, and so on, provided that landmarks can be located on the frames to be used. Because there is no correlation requirement on the frames, the enhancement result can be obtained not only from videos taken under very low speed, but also from the combined videos taken in different periods.

- **Practicality**

The general frame rate for a video camera is about 30 frames per second. In the basic experiments (see Chapter 5), 10 frames were used. In the applications to real videos (see Chapter 7), the maximum number of frames was 29. Significant improvements were achieved. Therefore, it is possible to recover the facial area once a one-second-long qualified video is captured.

8.2 Contributions

The main contribution of the work is to apply face models in facial image enhancement. In a broad sense, the proposed algorithm is also suitable for other areas involving in object restoration. As for face enhancement, if some frames for a particular object are obtained together with a proper object model, which is able to describe the main features of the object, is established, we can fit the model to each frame individually, transform the objects to the same shape by warping, and then frame average the warped results to recover the target frame. By warping the recovered target frame to each frame in the original sequence, the full sequence can be enhanced.

Expected applications exist in the areas of:

- medicine, where tumors or other diseases in patients are to be detected from the videos taken from endoscopies.
- remote sensing, where the earth sources or geographical mapping are tracked on videos acquired by satellites.

8.3 Future Work

As mentioned in Section 6.5, more work is required to reduce the ambiguity for landmark locating. This can be accomplished by defining landmarks in a more accurate way or introducing reference lines.

- **Defining landmarks in a more accurate way**

In the face model, a few landmarks are subject to the operator's subjective judgement. An example is the definition of the landmarks in the middle of the eyebrows

(landmark 1 and landmark 2, see Figure 3.2). Redefinition for such landmarks is necessary.

- **Introducing reference lines**

Because the human face is symmetrical, some reference lines will be helpful to locate the landmarks which depend on the others at a more accurate place. For example, the interface can ask input for the central line of the face first, then the system can create the eye line, nose line and mouth line automatically, once a eye corner, nose tip and a mouth corner are identified respectively. According to the definitions in Fig.3.2, all the landmarks related to these lines can be located easily and accurately.

In addition, the quantitative measurements between the degraded frames and the enhanced frames were performed in the experiments in Chapter 5. Extensive subjective testing on image qualities before and after enhancement is required as another item of future work.

In the long term, it may eventually be possible to automate the system, although the extreme degradation of video images in typical applications such as security, makes this a challenging problem.

References

[1] Robert Forchheimer and Torbjörn Kronander, "Image Coding — From Waveforms to Animation," in *IEEE Transactions on Acoustics*, Speech, and signal Processing, Vol.37, No.12, pp 2008-2023, December 1989.

[2] R. H. McMann *et al.*, "A Digital Noise Reducer for Encoded NTSC Signals," in *SMPTE J.*, vol.87, pp.129-133, Mar.1978.

[3] J. R. Sanders, "Fully Adaptive Noise Reduction for a Television Network," in *Television*, vol. 18, pp.29-33, May-June 1980.

[4] T. J. Dennis, "Nonlinear Temporal Filter for Television Picture Noise Reduction," in *IEE Proc.*, part G, vol.127, pp.52-56, Apr.1980.

[5] T. S. Huang and Y. P. Hsu, "Image Sequence Enhancement," in *Image Sequence Analysis*, T. S. Huang, Ed. Berlin, Germany: SpringerVerlag, 1981, pp.289-309.

[6] Eric Dubois and Shaker Sabri, "Noise Reduction in Image Sequences Using Motion-Compensated Temporal Filtering", in *IEEE Transactions on Communications*, Vol. Com-32, No.7, July 1984.

[7] Parke, F. I., "A parametric model for human faces," in *Tech. Report UTEC-CSc-75-047 Salt Lake City*: University of Utah, 1974.

[8] Parke, F.I., "Parameterized Models for Facial Animation," in *IEEE Computer Graphics and Applications*, 2(9), Nov. 1982, pp.61-68.

[9] Eaves J. and A.Paterson, "FACE – Facial Automated Composition and Editing," in *Ausgraph*'90:329-333.

[10] H. Delinguette, G. Subsol, S. Cotin, and J. Pignon, "A Craniofacial Surgery Simulation Testbed," in *Technical Report 2199*, INRIA, France, February 1994.

[11] F.I. Parke, "Computer Generated Animation of Faces," in *Master's Theses, University of Utah*, Salt Lake city, June 1972, UTEC-CSc-72-120.

[12] F.I. Parke, "Parameterized Models for Faical Animation," in *IEEE Computer Graphics and Application*, 2(9):61-68, 1982.

[13] S. M. Platt, "A system for Computer Simulation of the Human Face," in *Master's thesis*, The Moore School, Pennsylvania, 1980.

[14] S. M. Platt and N. I. Badler, "Animating Facial Expressions," in *Computer Graphics*, 15(3):245-252, 1981.

[15] K. Waters, "A muscle Model for Animating Three-Dimensional Facial Expressions," in *Computer Graphics* (SIGGRAPH'87), 21(4): 17-24, July 1987.

[16] D. Terzopoulos and K. Waters, "Analysis of Dynamic Facial Images Using Physical and Anatomical Models," in *Proc. Third International Conference Computer Vision* (Osaka, Japan), 1990, pp.306-331.

[17] K. Waters, "Modelling 3D Facial Expression" and " A Physical model of Facial Tissue and Muscle Articulation," in *SIGGRAPH'90 Tutorial*, State of the Art in Facial Animation, 1990.

[18] John A. Robinson, Jason Fischl and Bryan Miller, "Muscle-based Analysis/Synthesis Video Coding," in *Canadian Journal of Electrical and Computer Engineering*, Vol. 23, Nos 1-2, 1998.

[19] Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.19, no.7, pp.743-756, July 1997.

[20] D. E. Pearson and J. A. Robinson, "Visual Communication at Very Low Data Rates," in *Proc. IEEE 73*, pp.795-812, April 1985.

[21] M. Fischler and R. Elschlager, "the Representation and Matching of Pictorial Structures", in *IEEE Transactions on Computers*, 22(1), 1973.

[22] T. Beier and S. Neely. "Feature-based Image Metamorphosis". In *Computer Graphics*, vol 26(2), pp 35-42, New York, NY, July 1992. Proceedings of SIGGRAPH '92.

[23] K.-K. Sung, "Learning and Example Selection for Object and Pattern Detection," *AI Technical Report 1572*, MIT AI Lab, Jan. 1996.

[24] George Wolberg, "Digital Image Warping," *IEEE Computer Society Press,* Los Alamitos, California, 1992.

[25] Li-Te Cheng and John Robinson, "MCLGallery: A Framework for Multimedia Communications Research," in *Proceedings of CCECE'98*, May 1998, Waterloo.

[26] Jan Teuber, "Digital Image Processing", *Prentice Hall*, 1993.

[27] Anil K. Jain, "Fundamentals of Digital Image Processing", *Prentice Hall* , 1989.

[28] Umbaugh, S.E. (1998). "Computer Vision and Image Processing: A Pracitical Approach Using CVIP Tools", *Prentice Hall PTR*, Upper Saddle River, New Jersey 07458.

Appendix 1: Linear Regression

We consider the problem of fitting a set of N data points (x_i, y_i) (i = 1, 2, …, N) composed by the average grey levels for all rows or columns on a facial frame to a straight non-vertical line model:

$$y(x) = y(x; k, p) = kx + p \tag{1}$$

where k is the slope, giving the change in the y axis value for each unit change along x axis, and p is the intercept, the point at which the line crosses the y axis when x is zero.

This problem is often called *linear regression*, a terminology that originated in the social sciences. The regression line also makes the sum of the squared residuals a minimum. Hence it is also called the "least squares line".

We assume that the summation of square error is a function with respect to k and p:

$$f(k, p) = \sum_{i=1}^{N} (y_i' - y_i)^2 = \sum_{i=1}^{N} (ki + p - y_i)^2 \tag{2}$$

Equation (2) is minimised to determine k and p. At its minimum, the derivatives of $f(k, p)$ with respect to k and p vanish.

$$\begin{cases} \dfrac{\partial f(k, p)}{\partial k} = 0 \\ \dfrac{\partial f(k, p)}{\partial p} = 0 \end{cases} \tag{3}$$

Therefore, we get equation (4) by substituting (2) to (3):

$$\begin{cases} \dfrac{\partial f(k,p)}{\partial k} = 2\sum_{i=1}^{N}(kx + p - y_i)i = 0 \\ \dfrac{\partial f(k,p)}{\partial p} = 2\sum_{i=1}^{N}(kx + p - y_i) = 0 \end{cases} \tag{4}$$

These conditions can be rewritten in a convenient form if we define the following sums,

$$\begin{aligned} S_i &= \sum_{i=1}^{N} i \\ S_{ii} &= \sum_{i=1}^{N} i^2 \\ S_y &= \sum_{i=1}^{N} y_i \\ S_{yi} &= \sum_{i=1}^{N} y_i i \end{aligned} \tag{5}$$

The solution of these two equations in two unknowns is calculated as:

$$\begin{cases} p = \dfrac{S_{yi}S_i - S_{ii}S_y}{S_i^{\,2} - NS_{ii}} \\ k = \dfrac{S_y S_i - NS_{yi}}{S_i^{\,2} - NS_{ii}} \end{cases} \tag{6}$$

Equation (6) gives the solution for the best-fit model parameters k and p. An example of the regression lines for the shaded facial image in Fig.3.10(a) are depicted previously in Fig.3.11.

Appendix 2: Composing a Plane by Two Lines

Fig.1 shows a plane composed by two given lines which intersect at (0, 0, p).

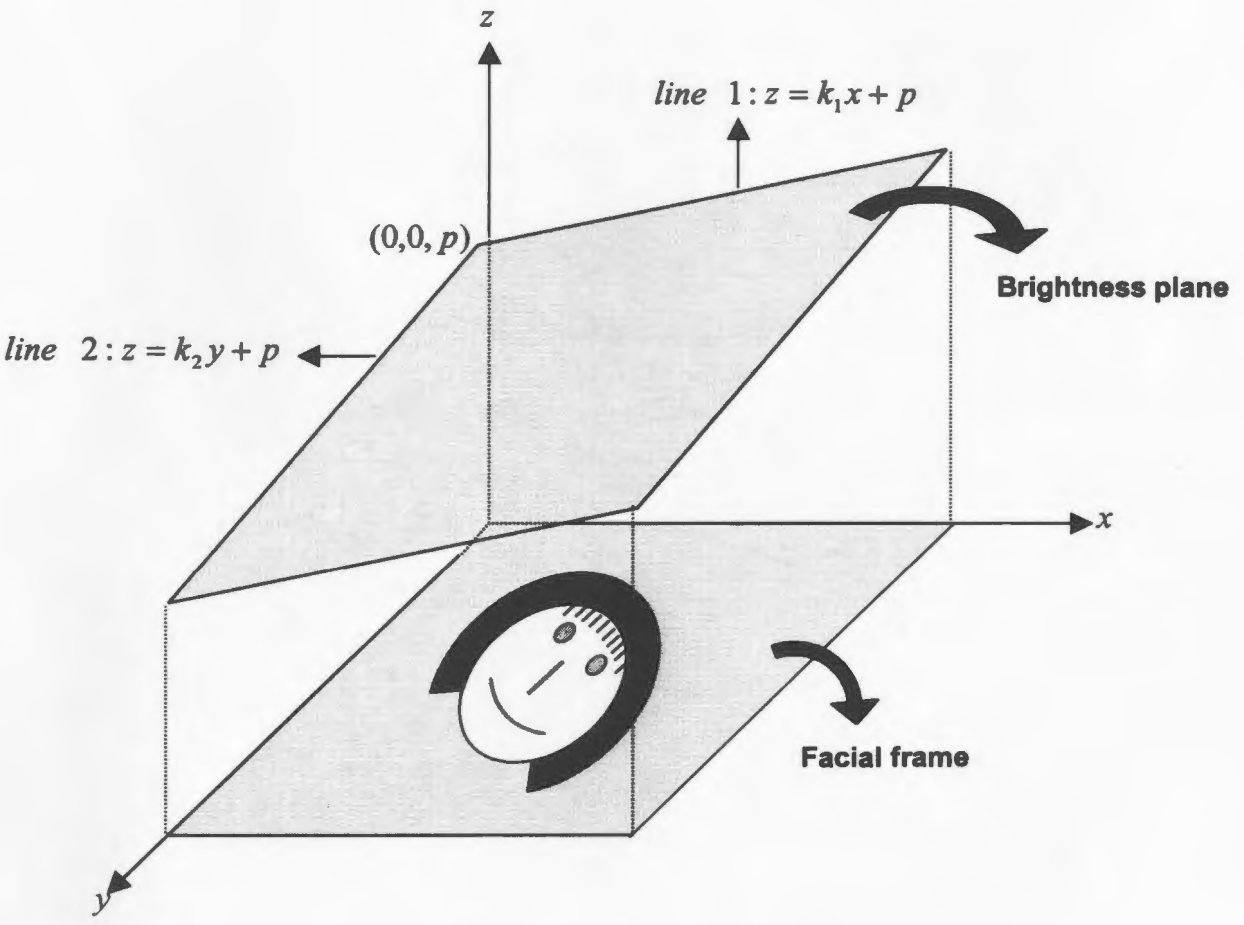


Fig.1 the Plane Composed by Two Lines

The line equations are:

$$line\ 1: \begin{cases} y = 0 \\ z = k_1 x + p \end{cases}$$

$$line\ \ 2: \begin{cases} x = 0 \\ z = k_2 y + p \end{cases}$$

The plane equation will be calculated in this appendix.

Suppose the plane equation is $Ax + By + Cz = 1$ with unknown parameters A, B and C. From Fig.1, we can easily find three points lying on the plane:

$(0, 0, p)$	intersection of two lines;
$(0, 1, k_2 + p)$	from line 2;
$(0, 1, k_1 + p)$	from line 1;

Three equations are immediately obtained by substituting co-ordinates of these points:

$$\begin{cases} C \times p = 1 \\ B \times 1 + C \times (k_2 + p) = 1 \\ A \times 1 + C \times (k_1 + p) = 1 \end{cases}$$

If p is not equal to 0, the parameters for the plane equation are:

$$\begin{cases} C = \dfrac{1}{p} \\ B = -\dfrac{k_2}{p} \\ A = -\dfrac{k_1}{p} \end{cases}$$

Therefore, the plane determined by the two lines can be represented by the equation:

$$z = k_1 x + k_2 y + p$$

This equation is also suitable for the situation when p is equal to 0.

Appendix 3: Application for Permission

In this appendix, the documents, *the Certification of Ethical Acceptability for Research Involving Human Subjects* and *Consent Form for Participation in Facial Landmarks Experiment*, used for subjective tests application in this thesis are attached.

Appendix 3.1 :

Certification of Ethical Acceptability for Research Involving Human Subjects

Date:	Nov. 21, 1997
Name of Applicant:	Xiaomeng Ping
Department:	Engineering & Applied Science
Agency:	N/A
Title of Project:	Supervised Object-Based Temporal Filtering for Enhancement of Moving Facial Images

1. Name(s) of principal investigator(s):

   Xiaomeng Ping

2. Name of supervisor (if applicable):

   Dr. John A. Robinson

3. Title of investigation:

   Experiment on Locating Facial Landmarks by Different Operators

4. Duration of investigation:

   The research will be conducted from approximately *Nov. 1997* to *Feb. 1998* (insert dates).

5. Classification of research:

   a. Faculty funded research.
   b. Faculty non-funded research.
   c. Ph.D thesis research
   × d. M.Sc thesis research.
   e. Honours thesis research.
   f. Student research conducted as part of a course.
   g. Other (please specify).

6. Medical procedures:

Does the research use medical procedures? No.

7. Other institutions or agencies: N/A

8. Brief description of research:

We are investigating an object-based system for recovering accurate facial images from noisy and blurred videos. A human-supervised model-fitting procedure is used in our system. That is, the face model is composed from a set of landmarks which describe the main facial feature points, and these landmarks are overlaid, frame-by-frame on the video, by a human. This procedure is the foundation of our system.

Obviously, different sets of landmarks on the same frame will be generated by different operators. A supervised system will be unreliable and impractical if variation between operators causes significantly different results. Therefore, it is necessary to test whether our supervised system is sensitive to the landmark variations created by different operators.

9. Description of research procedures:

In the experiment, 10 subjects are needed. Given several groups of frame sequence (up to 4 sequences of 10 frames), the subjects are required to identify the landmarks on each frame manually and save the locations of the landmarks in a data file in the corresponding directory. The work will be done using our *Facial Landmark Locating Interface*, which is an easy-to-use executable program running in Windows 95. The working environment, including selecting frame sequences, assigning the individual directory and instructing on the user interface are prepared by the experimenter. The time to complete the experiment will be recorded.

The equipment in the experiment is a computer installed with our user interface. We have no questionnaires or other stimuli in the experiment. All actions for the subjects are similar to conventional computer operations, such as clicking on buttons of the mouse. There are no risks associated with the experiment.

The data files generated by each subject will be used to recover de-noised facial frames. By comparing the facial frame recovered by each subject, we can conclude if

our supervised system is feasible for enhancement of facial image sequence and independent of the operator.

10. Research Proposal: N/A

10. Investigator's statement of ethical acceptability:

I hereby certify that the research summarized on this form and in the supporting documents deals with the human subjects in an ethical manner. I further certify that any substantial changes in procedures bearing upon ethical matters will be submitted to the Ethics Committee for approval.

Signature of investigator(s): _____

Signature of Supervisor (if applicable): _____

11. Today's date: Nov.21, 1997

12. Judgment by Ethics Committee:

We, the undersigned members of the Faculty of Science Ethics Committee, believe the research summarized on this form and in the supporting documents deals with the human subjects in an ethical manner.

Name	Faculty/Department	Position	Field of Research	Signature

Appendix 3.2:

Consent Form for Participation in Facial Landmarks Experiment

1. Name(s) of principal investigator(s):

Student: Xiaomeng Ping

Supervisor: Dr. John A. Robinson

2. Sponsors of the research:

National Sciences and Engineering Research Council of Canada (NSERC)

Newtel Communications (Newtel)

Northern Telecom (Nortel)

3. Title of investigation:

Experiment on Locating Facial Landmarks by Different Operators

4. Purpose of the investigation:

The proposed investigation is for our current research — *Supervised Object-Based Temporal Filtering for Enhancement of Moving Facial Images.*

5. Purpose of the research:

In our research, we are investigating a system for recovering accurate facial images from noisy and blurred videos. A human-supervised procedure called *facial landmarks locating* is used in our system. That is, a set of landmarks which describe the main facial feature points (such as the tip of the nose and corners of the mouth) are overlaid, frame-by-frame on the video, by a human operator. This procedure is the foundation of our system.

Obviously, different sets of landmarks will be generated by different operators. A supervised system will be unreliable and impractical if variation between operators causes significantly different results. Therefore, we want to find

out whether our supervised system is sensitive to the landmark variations created by different operators.

6. Description of the investigation:

The equipment in the investigation is a computer installed with our user interface, an easy-to-use executable program running in Windows 95. We have no questionnaires or other stimuli in the experiment. All actions for the subjects are similar to conventional computer operations, such as clicking on buttons of the mouse.

No medical procedures are used in the investigation. There are no potential harms and inconveniences associated with the investigation.

The data files generated by subjects will be collected for our research. To ensure confidentiality, the subject's name will not be attached to the corresponding data file.

It will take approximately 3 hours for each subject's participation. All the participation in the investigation is voluntary. You have the right to refuse and the right to withdraw at any time without prejudice.

7. Statement:

I consent to participate in the experiment.

Name:_____ Date:_____

Appendix 4:

Instructions on Facial Landmark Locating System

Experiment Description

In the experiment, there are altogether 4 groups of frame sequences which are stored in directories `D:\fl\sequence_1\`, `D:\fl\sequence_2\`, `D:\fl\sequence_3\` and `D:\fl\sequence_4\` respectively. Each frame sequence consists of 10 facial images. In addition, a directory `D:\fl\sequence_0\` is provided with two images for you to practice on. You are required to locate landmarks on all these frames and save the landmark locations in your own directory given as `D:\fl\Your_name\sequence_number\`. The locating work should be done in the sequence from sequence_1 to sequence _4. The procedure of landmark locating is as follows:

Step 1: Starting the system

- The *Facial Landmark Locating Interface* will be displayed on the screen throughout the experiment.
- In the interface, the face model consists of a set of landmarks indicated by red dots. The eye line, nose line, mouse line and facial vertical medium line are also shown for reading convenience. The detailed definitions of landmarks are attached *in Appendix I* at the end of the instructions. All your operations are performed in *Source Bitmap Control* group and on the *Input Image* window beside it.

Step 2: Locating landmarks

Load Facial Images:

- In *Source Bitmap Control* window, Click **Load Images**. A dialog box will be opened.

- Select a frame (`D:\fl\sequence_number\frame_number.bmp`) from one of frame sequences in the opened file dialog.
- Click **Ok**, then the frame will be loaded into picture box.

Start Locating:

- In *Source Bitmap Control* window, Click **Start Locating**.
- Each landmark is created by clicking the left button of the mouse on the frame. Because all the landmarks should be located in the appropriate sequence, the landmarks in the face model will blink in turn to indicate the next landmark to be located during the operating process. Watch the current blinking landmark and click left button of the mouse on the corresponding position of the loaded image.
- If you make a mistake, it can be corrected by clicking the right button of the mouse on the exact position of the incorrect landmark to prompt *Re-edit Message Box*. Click **Yes** if you want to re-edit the location and **No** if you want to delete it.
- If the loaded image is too noisy or heavily blurred to decide the landmark position, please locate the landmark on a reasonable location.
- It is recommended that you zoom the picture to make it larger before beginning locating work starts. You cannot zoom the picture once you have begun locating.

❖ **GENERAL ADVICE** ❖

Please aim for accuracy. The experiment will be timed, but it is more important to be accurate than to be fast.

Step 3: Save Locations of Landmarks

- When all the landmarks are located, a dialog box will be opened.
- Click **Save Locations** to save the positions of landmarks to a data file in your own directory **`D:\fl\your_name\sequence_number\file_number.dat`**. Note that the data file number should be the same as the frame number of the loaded image. For example, the locations for file `D:\fl\sequence_1\0.bmp` should be saved as `D:\fl\your_name\sequence_1\0.dat`. *Appendix 2* gives the convention on loaded image and the directory to save its corresponding data file.

❖ **WARNING** ❖

DON'T FORGET TO SAVE YOUR LOCATING WORK!

If you have any other questions, please feel free to ask Ping. Thank you for your help.

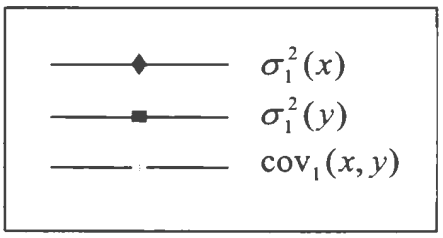
Appendix 5:

Variation on Landmark Locations by Subjects

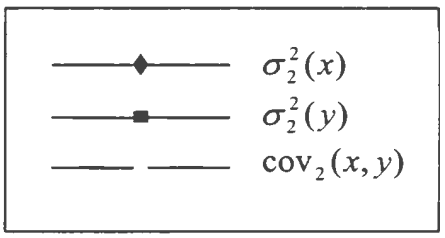
In this appendix, variance and covariance for landmarks in a particular testing frame are provided. Charts in *Appendix 5.1* are for $\sigma_1^2(x), \sigma_1^2(y)$ and $\mathrm{cov}_1(x,y)$. Charts in *Appendix 5.2* are for $\sigma_2^2(x)$, $\sigma_2^2(y)$ and $\mathrm{cov}_2(x,y)$. Frames indexed 1 to 10 are for sequence 1, indexed 11 to 20 for sequence 2, indexed 21 to 30 for sequence 3, and indexed 31 to 40 for sequence 4. Please refer to Section 6.5 for detailed definition of the parameters.

The legends for the charts in Appendix 5.1 and Appendix 5.2 are listed below.

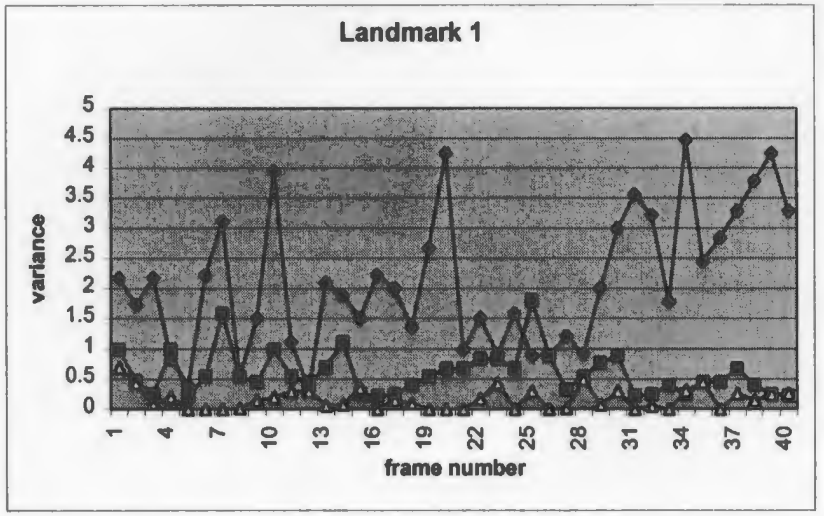
Legends for Appendix 5.1

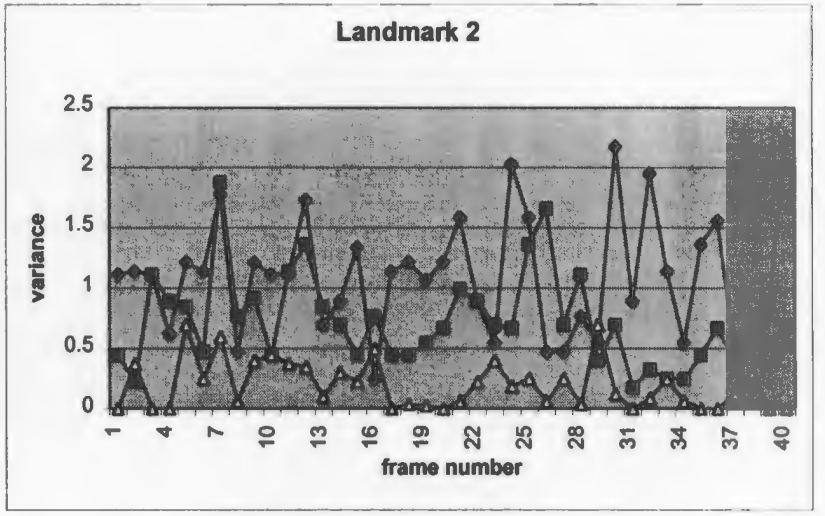


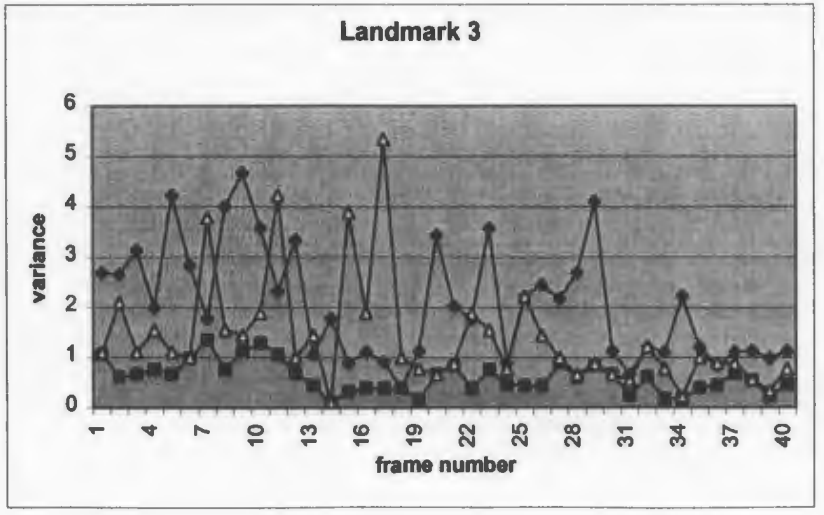
Legends for Appendix 5.2

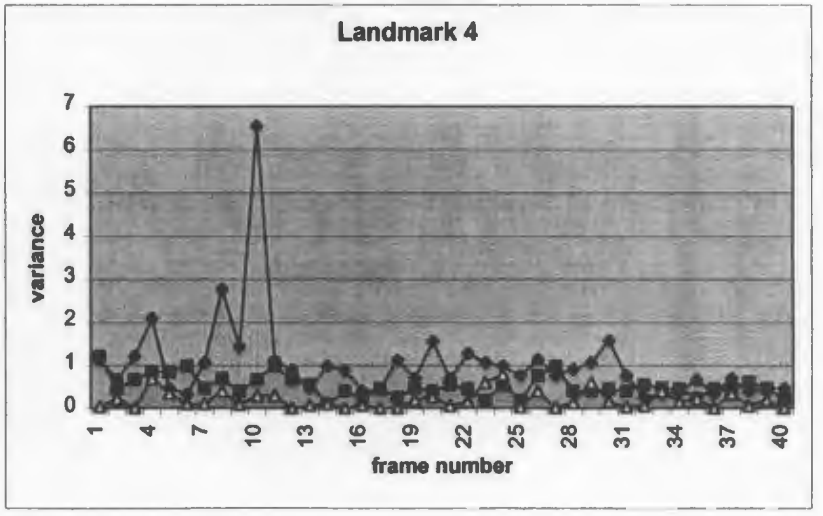


Appendix 5.1 Landmark Location Variation (for Landmarks 1 to 4)

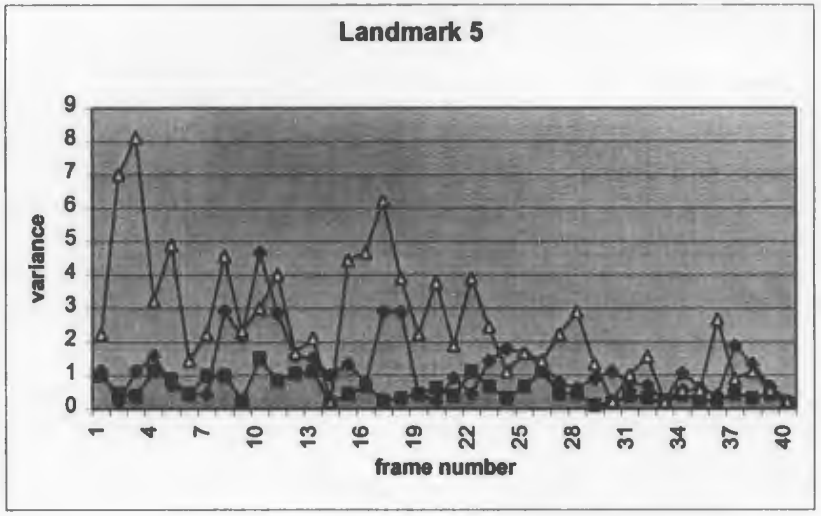


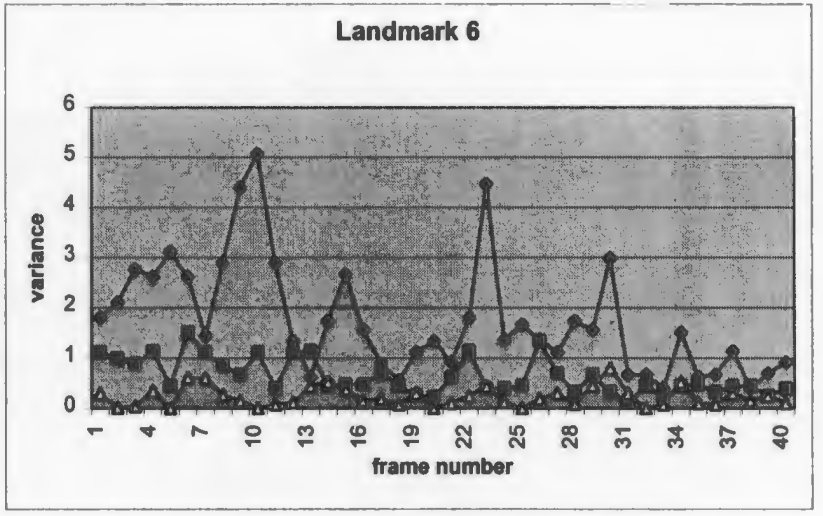


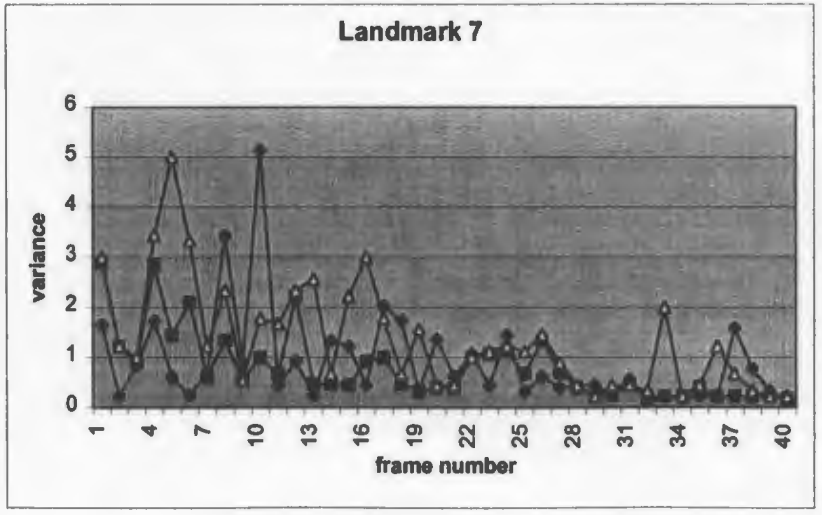


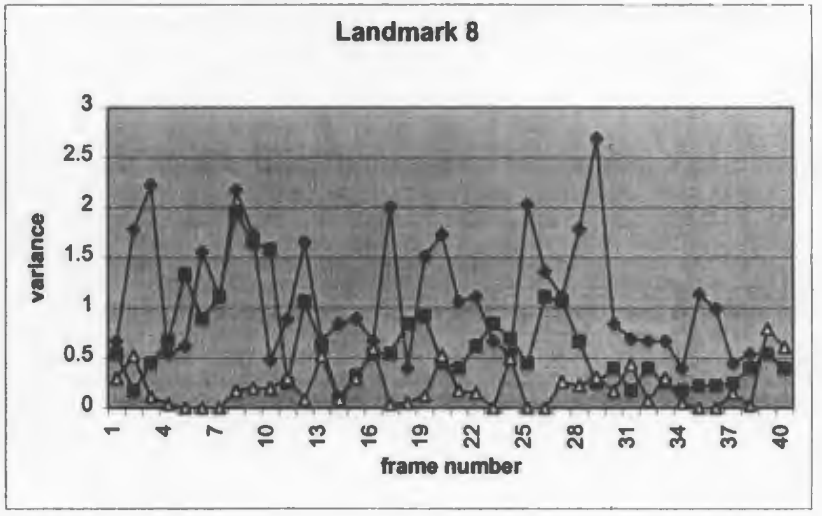


Appendix 5.1 Landmark Location Variation (for Landmarks 5 to 8)

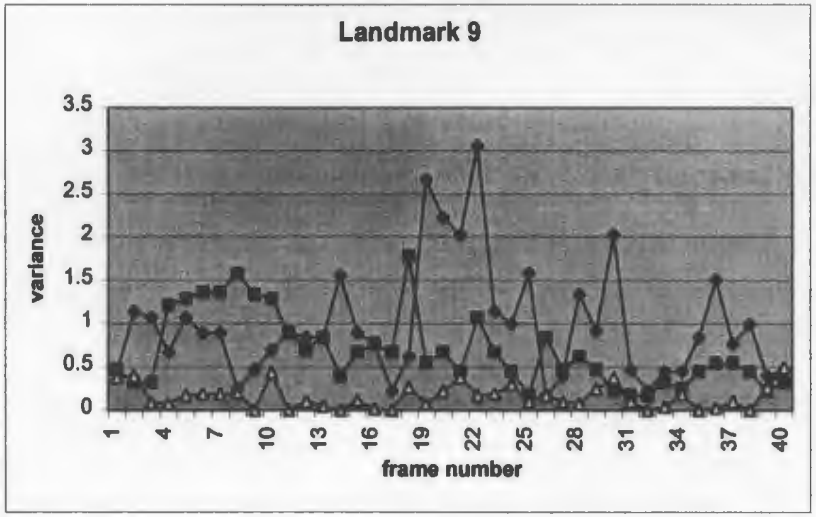


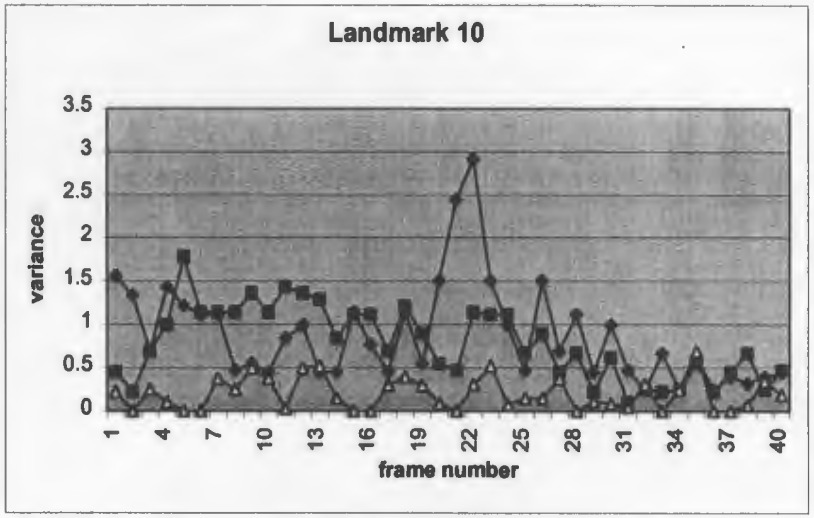


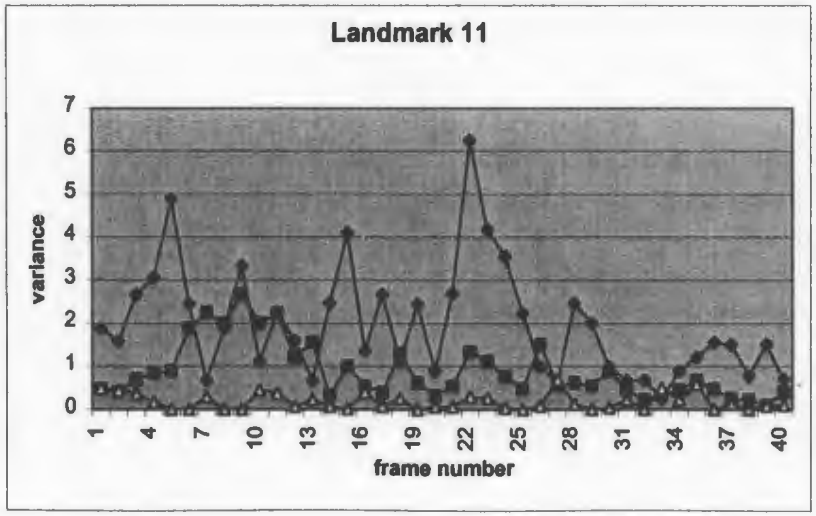


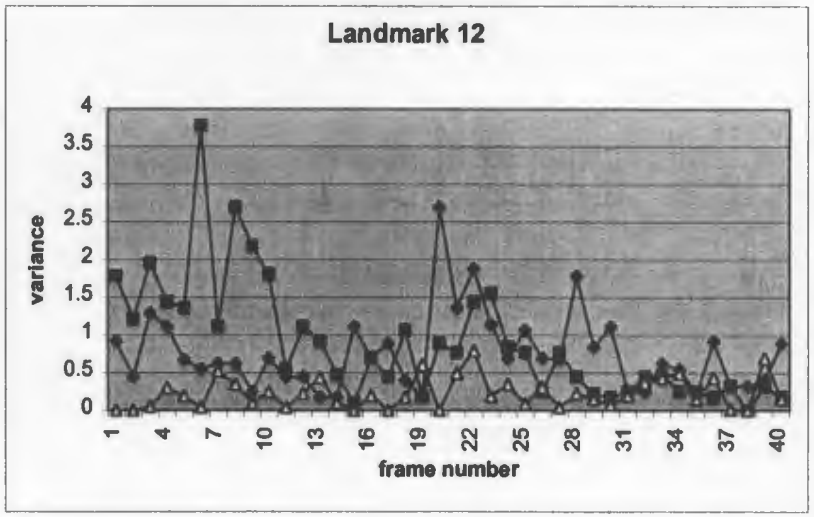


Appendix 5.1 Landmark Location Variation (for Landmarks 9 to 12)

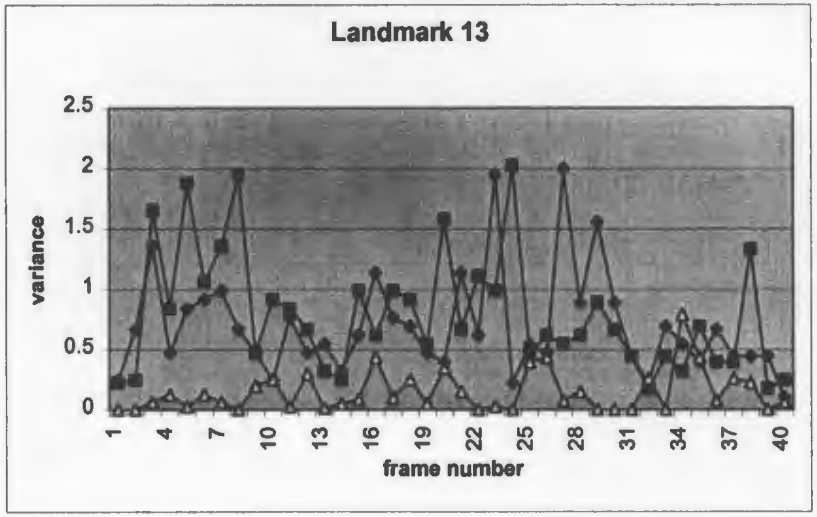


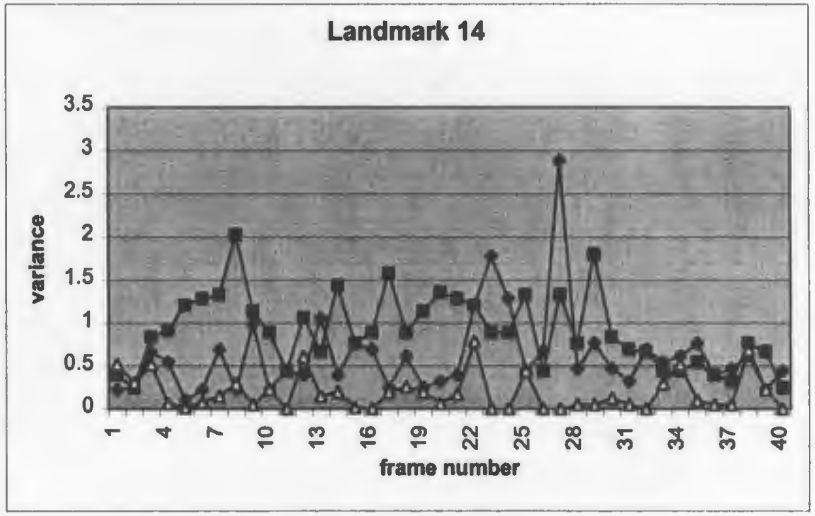


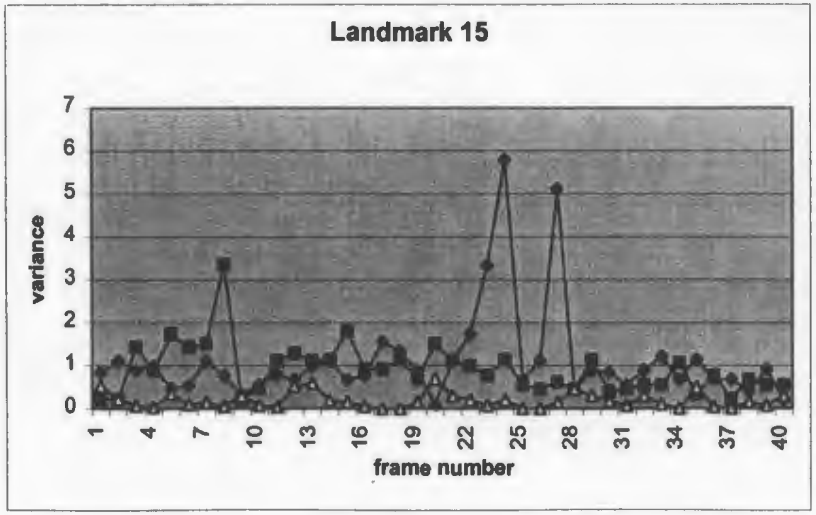


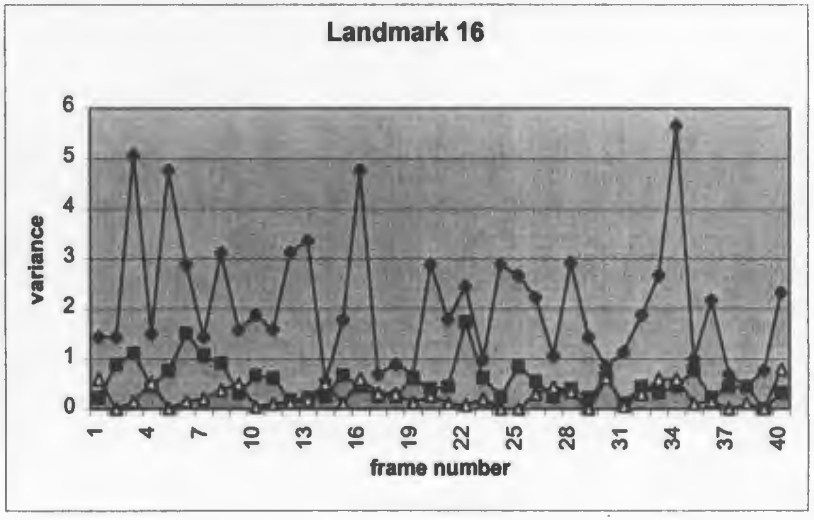


Appendix 5.1 Landmark Location Variation (for Landmarks 13 to 15)

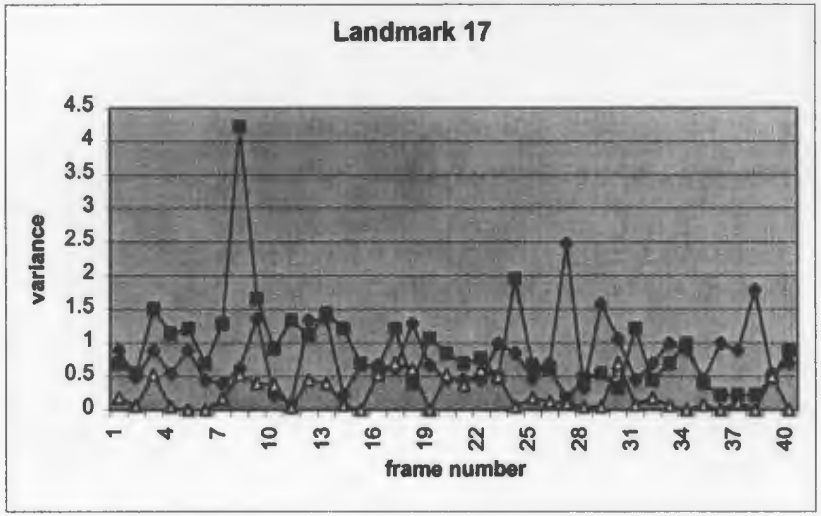


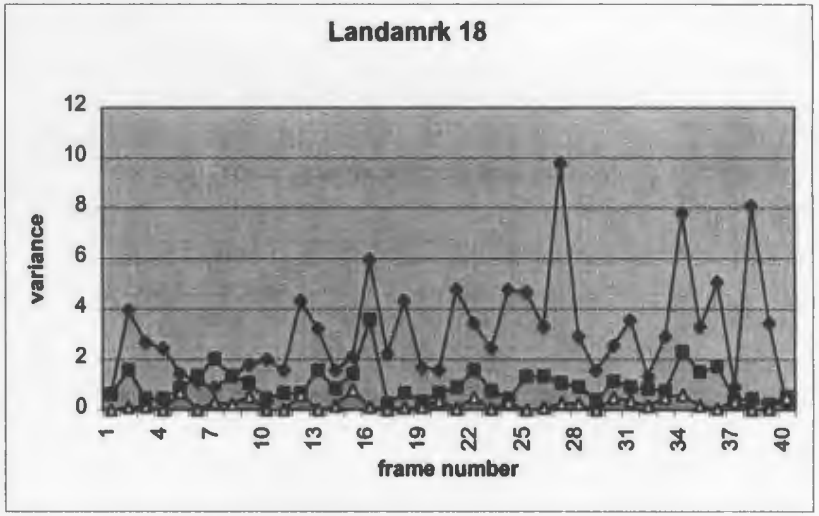


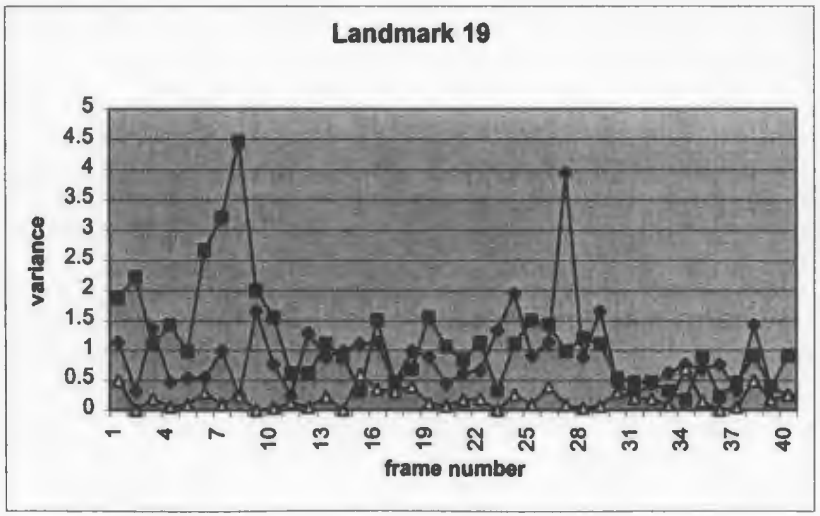


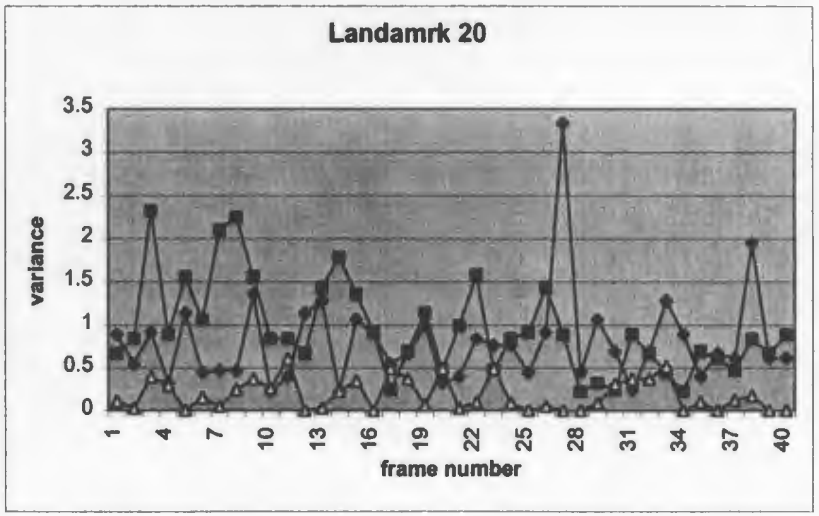


Appendix 5.1 Landmark Location Variation (for Landmarks 17 to 20)

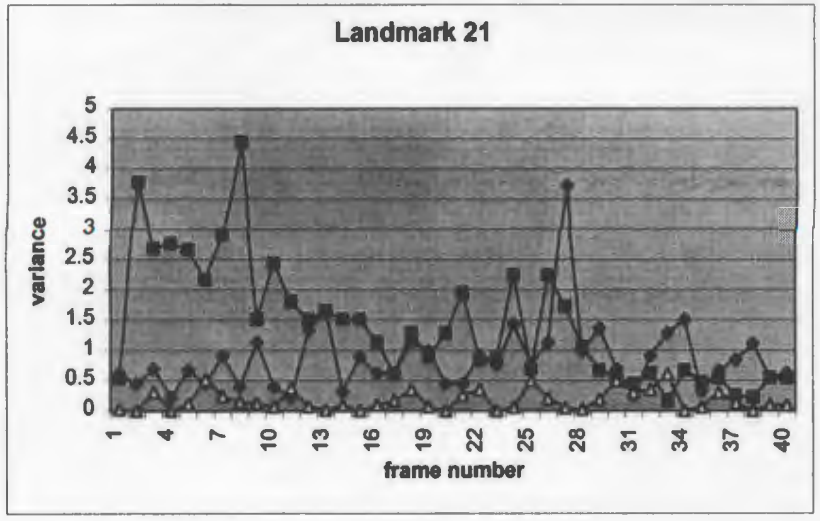


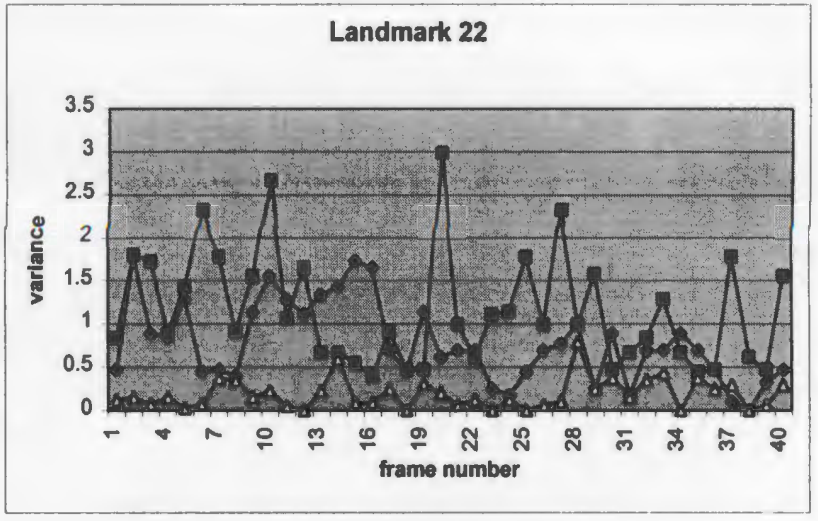


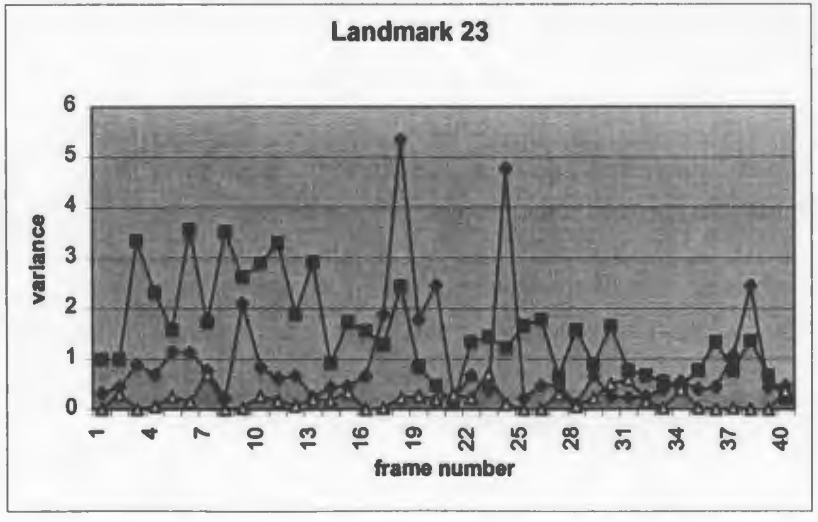


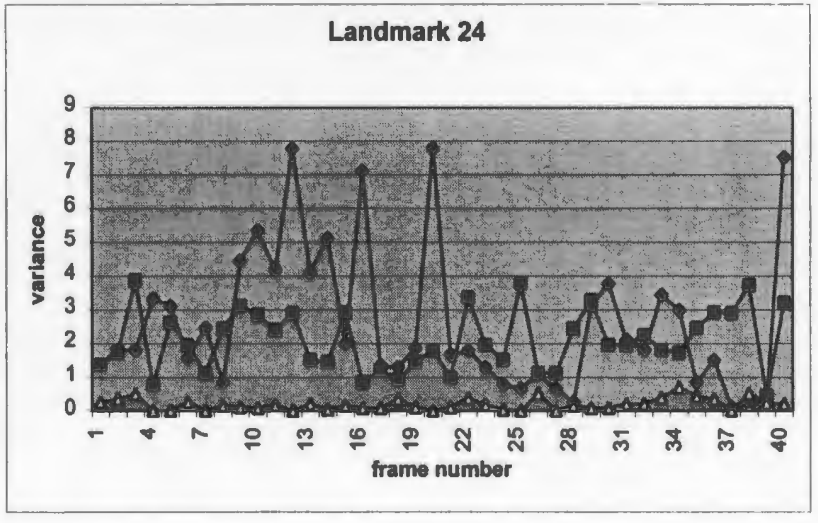


Appendix 5.1 Landmark Location Variation (for Landmarks 21 to 24)

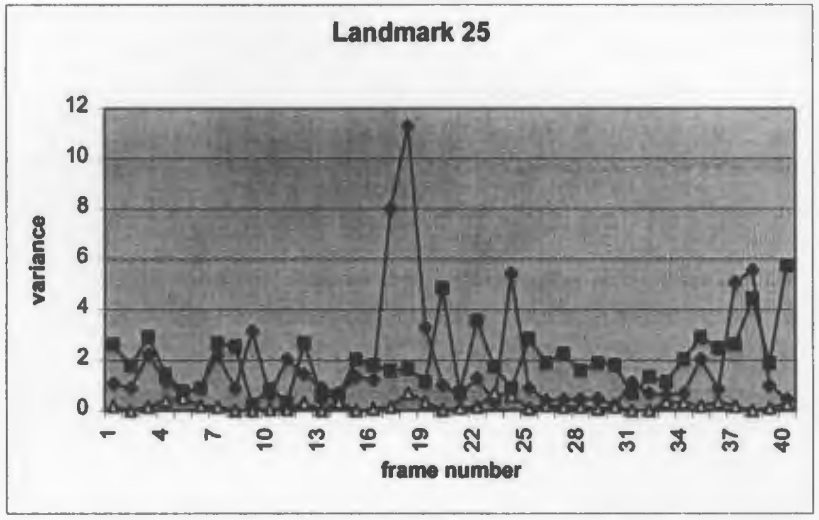


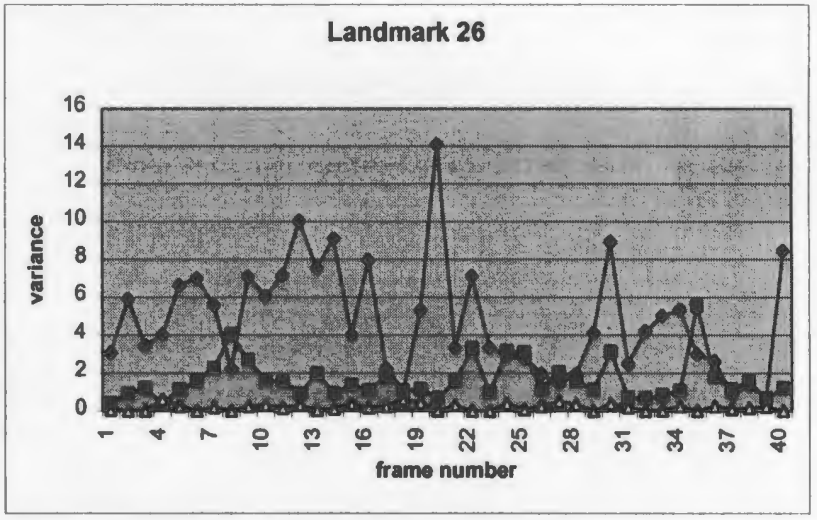


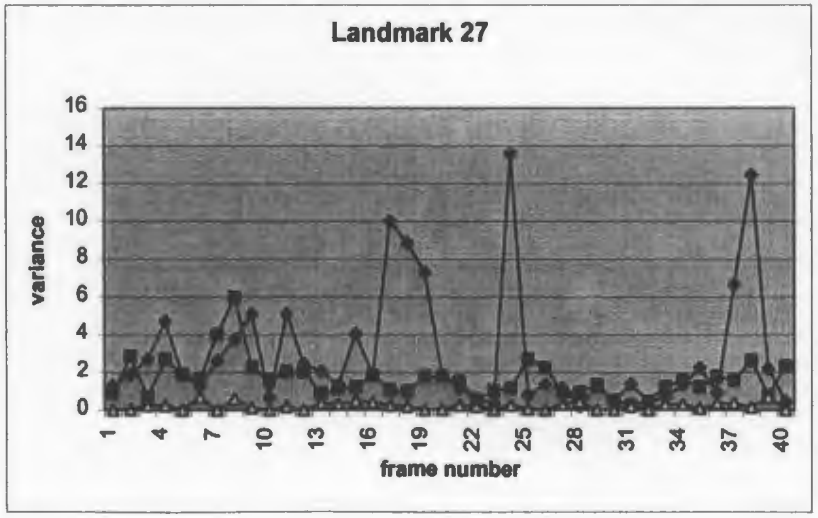


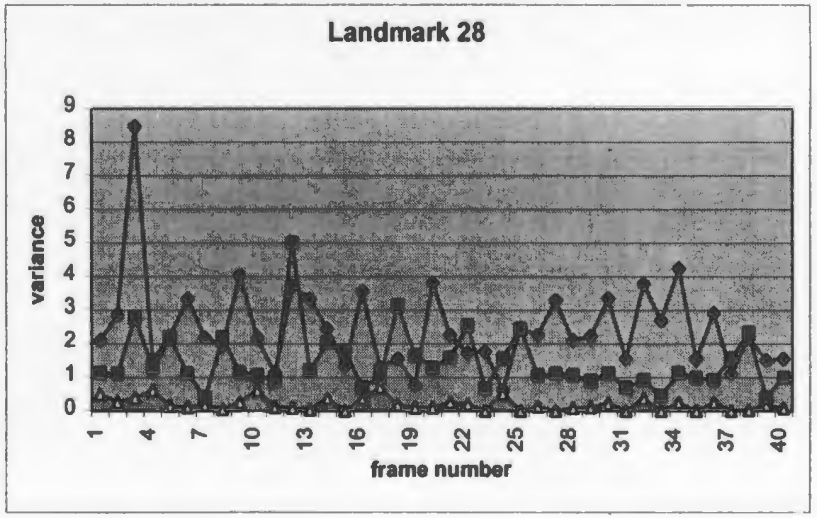


Appendix 5.1 Landmark Location Variation (for Landmarks 25 to 28)

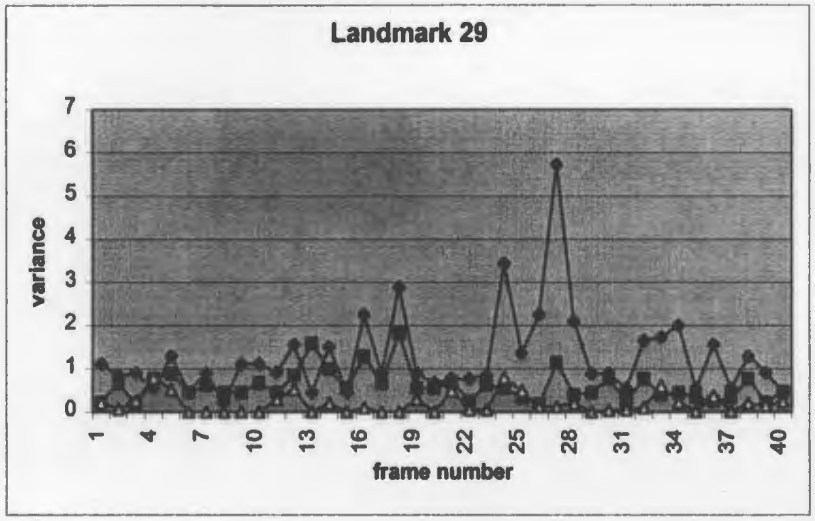


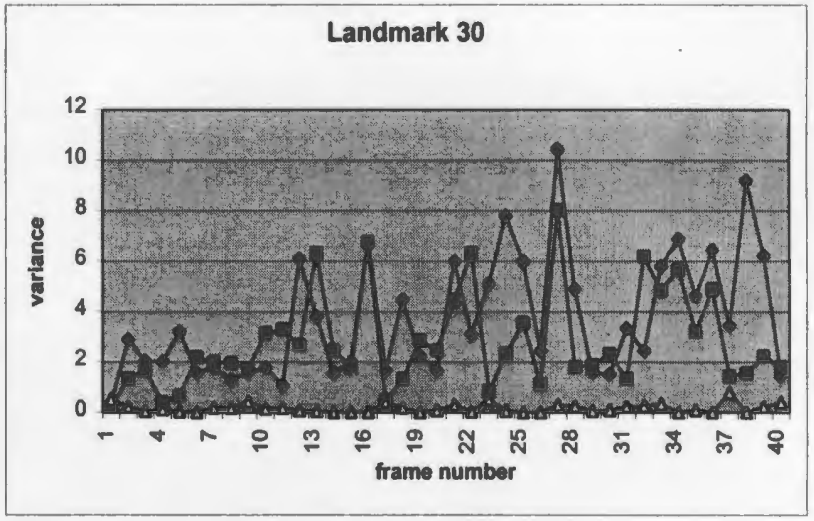




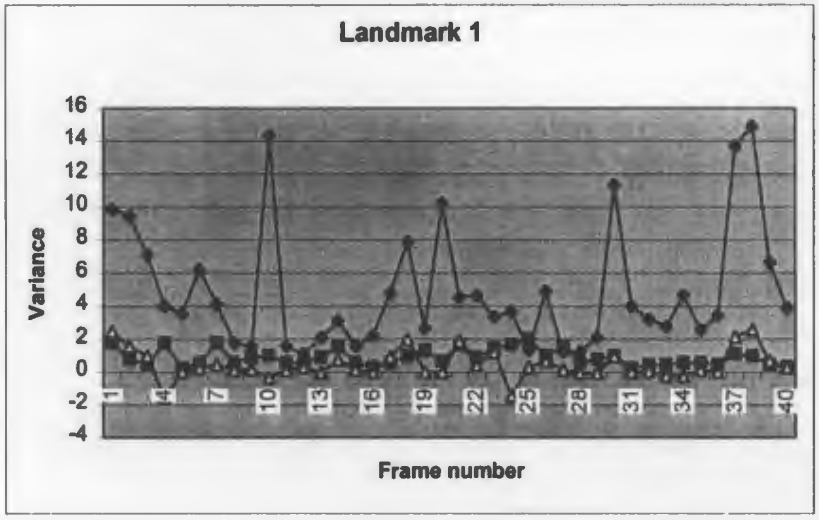


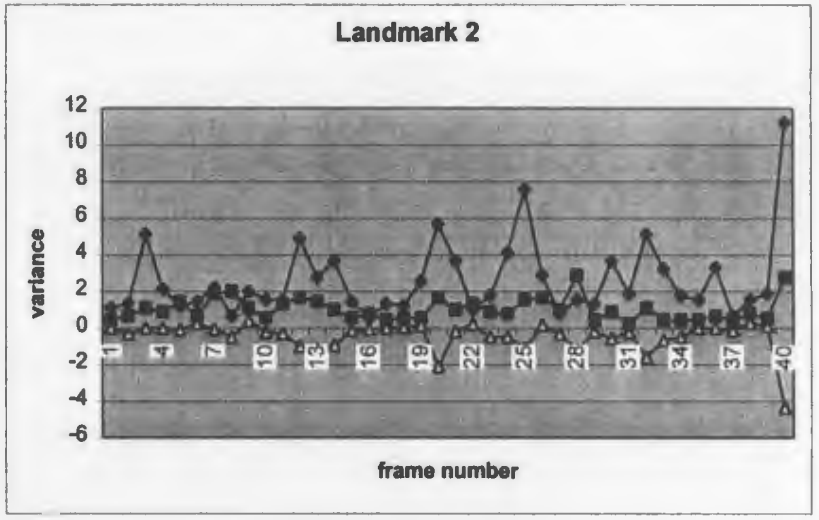
Appendix 5.1 Landmark Location Variation (for Landmarks 29 to 30)

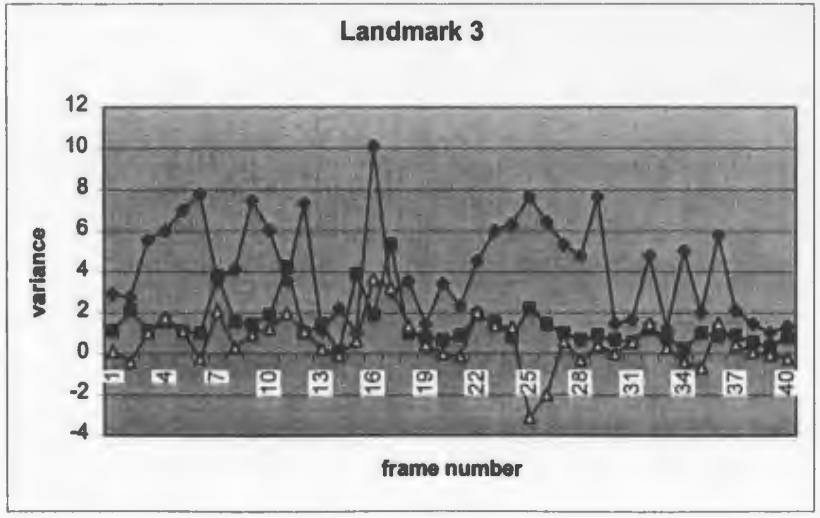


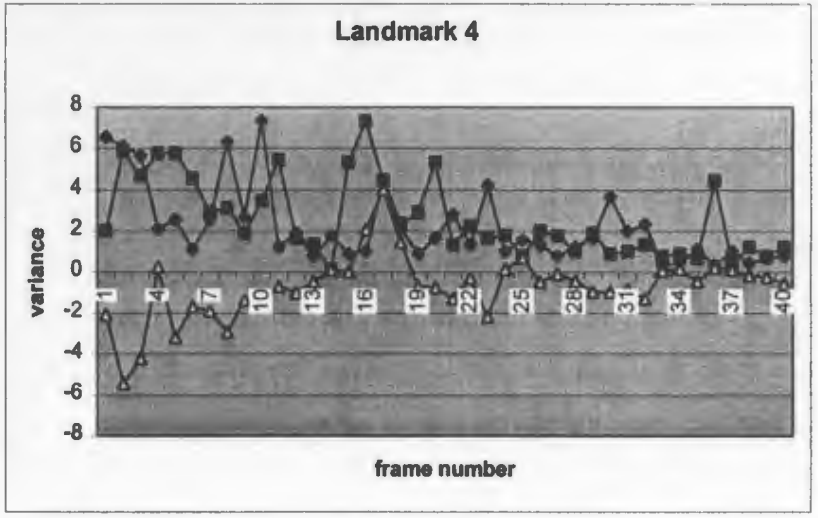


Appendix 5.2 Landmark Location Variation (for Landmarks 1 to 4)

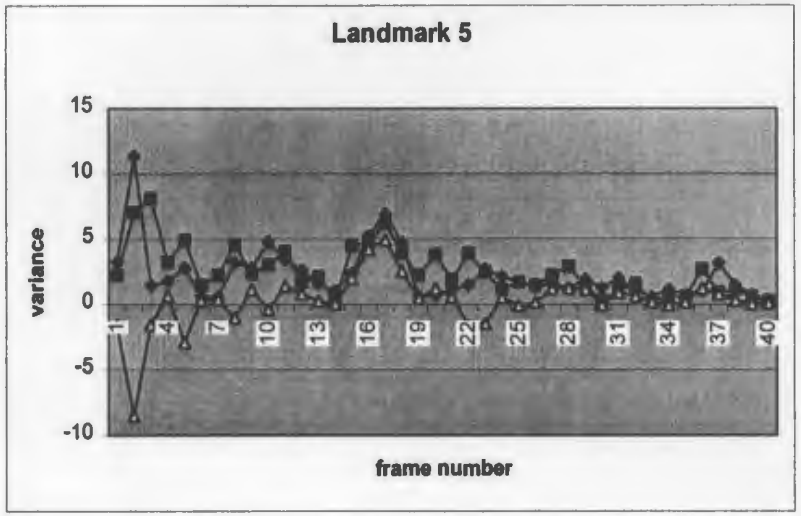


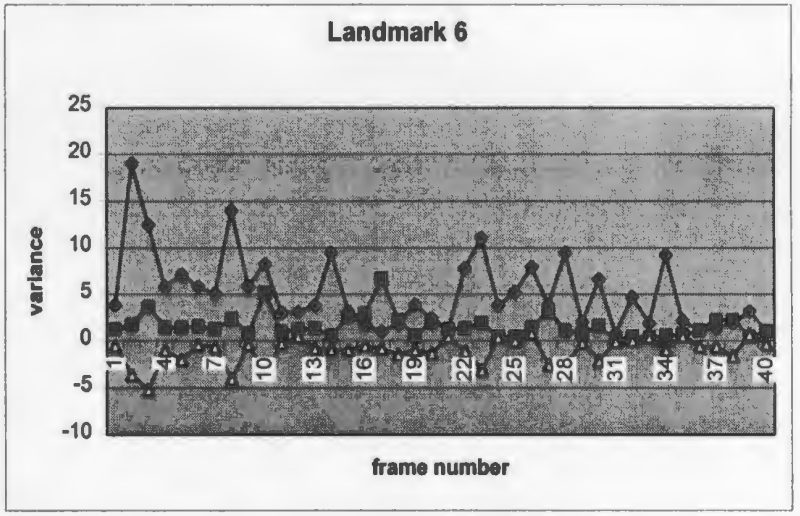


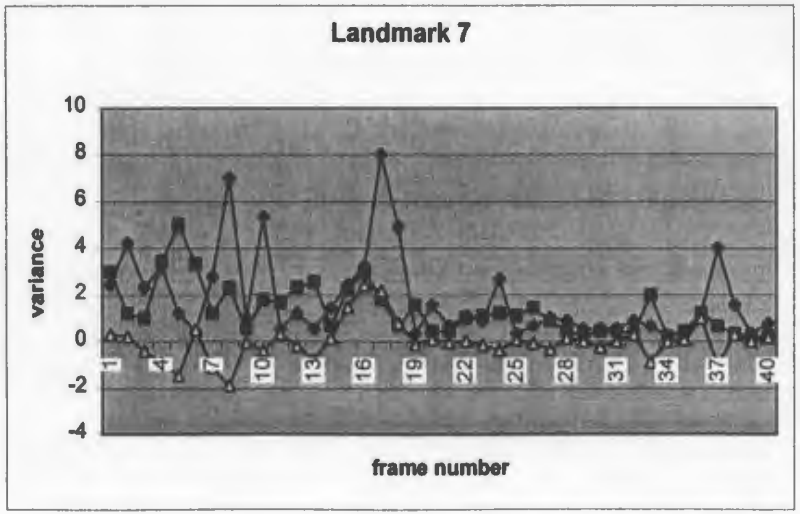


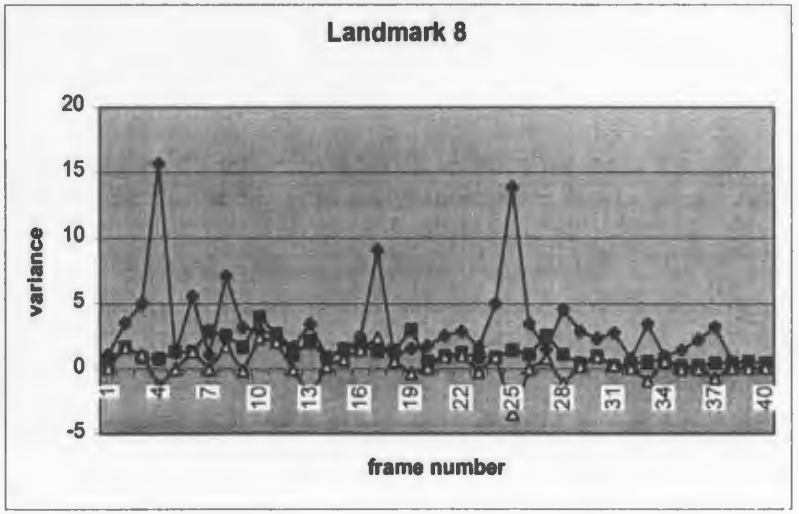


Appendix 5.2 Landmark Location Variation (for Landmarks 5 to 8)

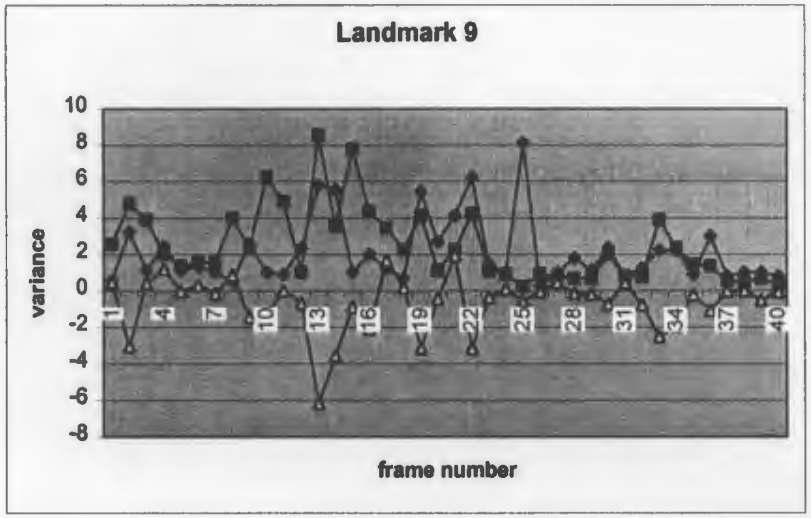


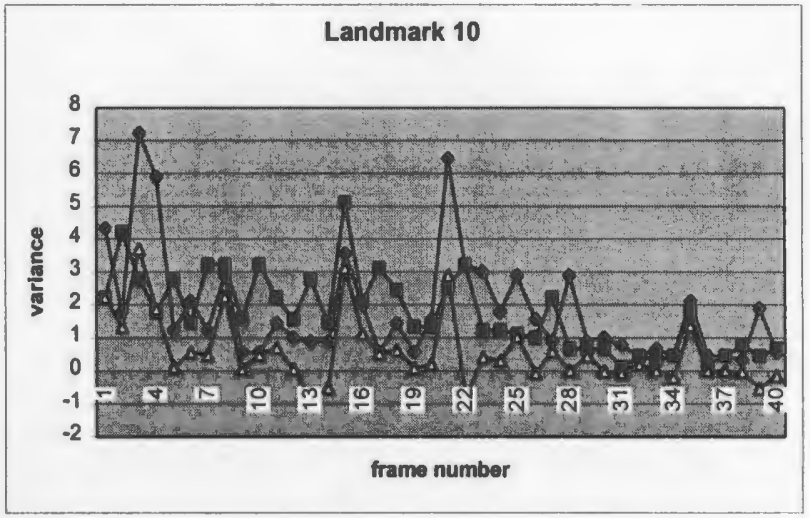


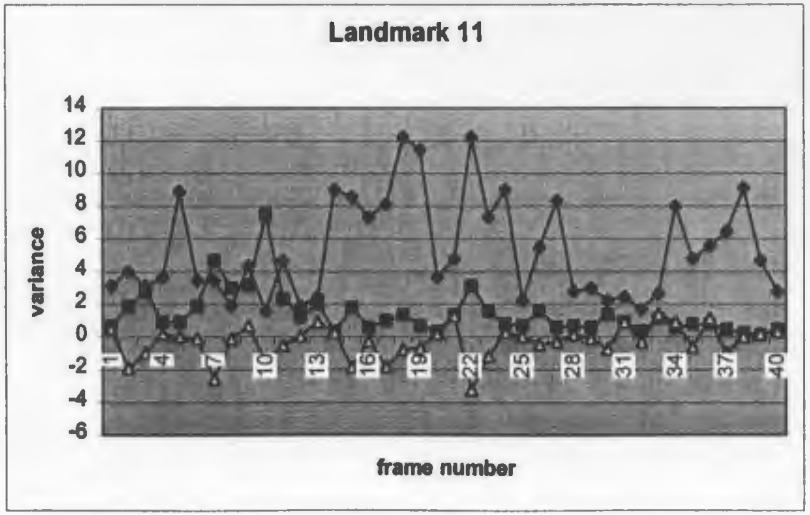


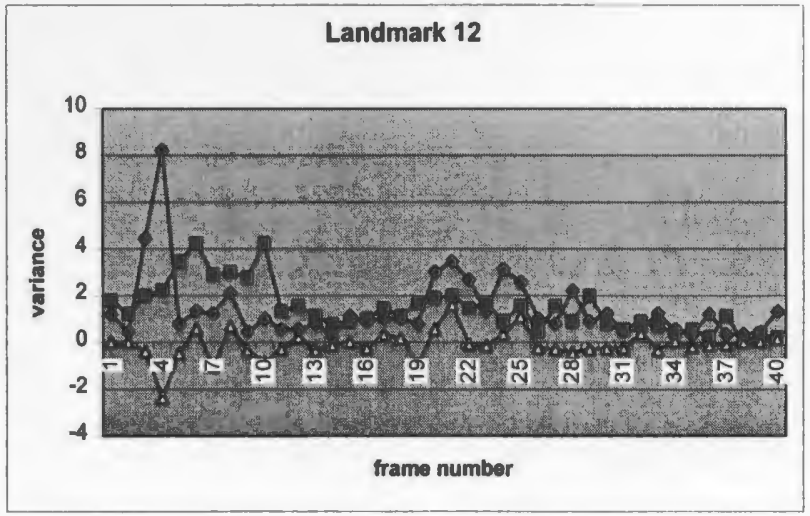


Appendix 5.2 Landmark Location Variation (for Landmarks 9 to 12)

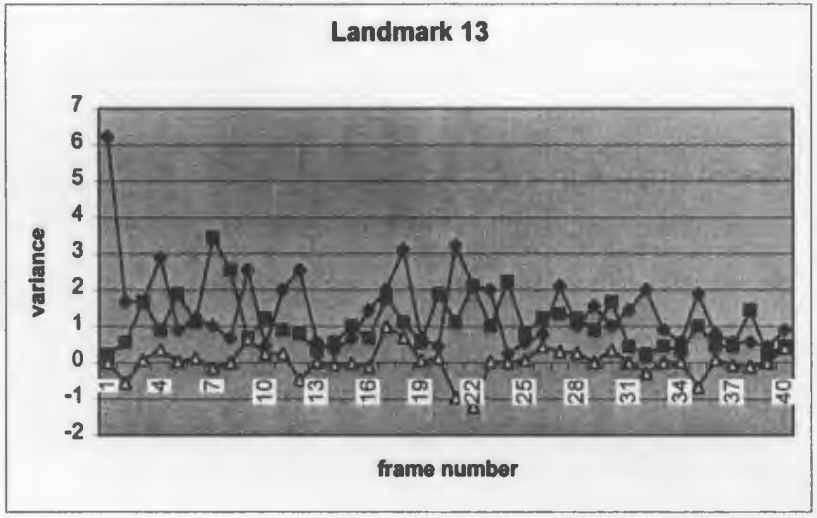


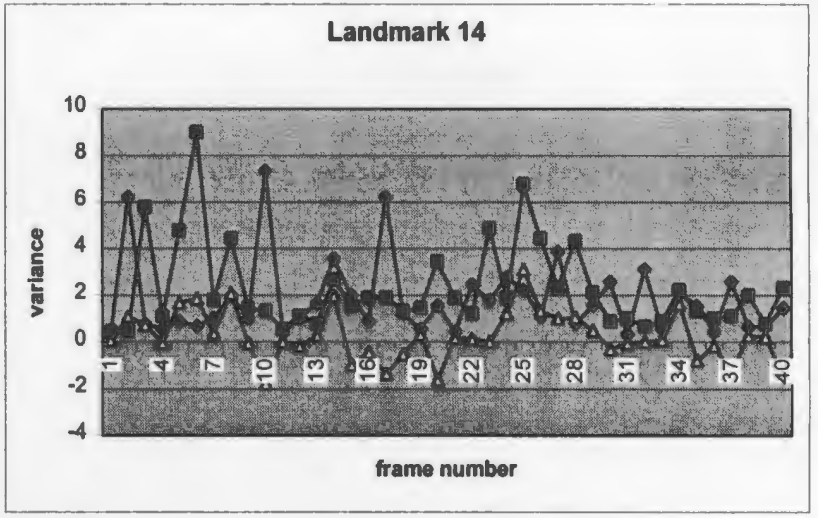


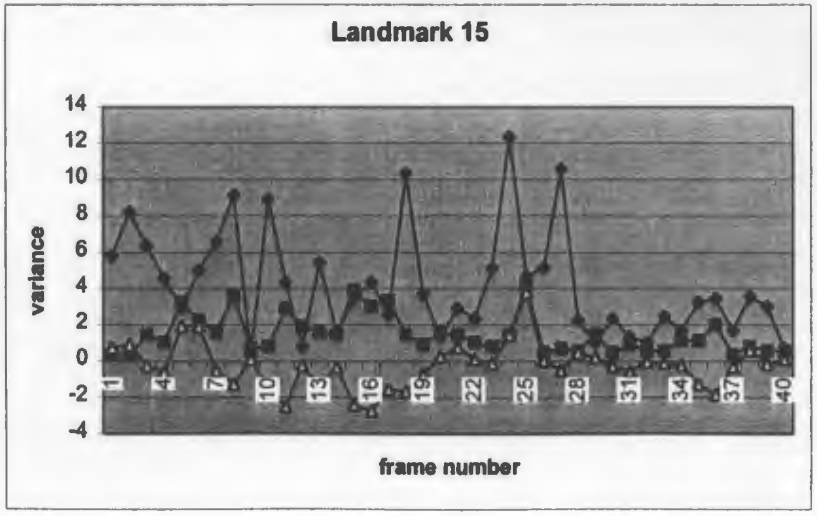


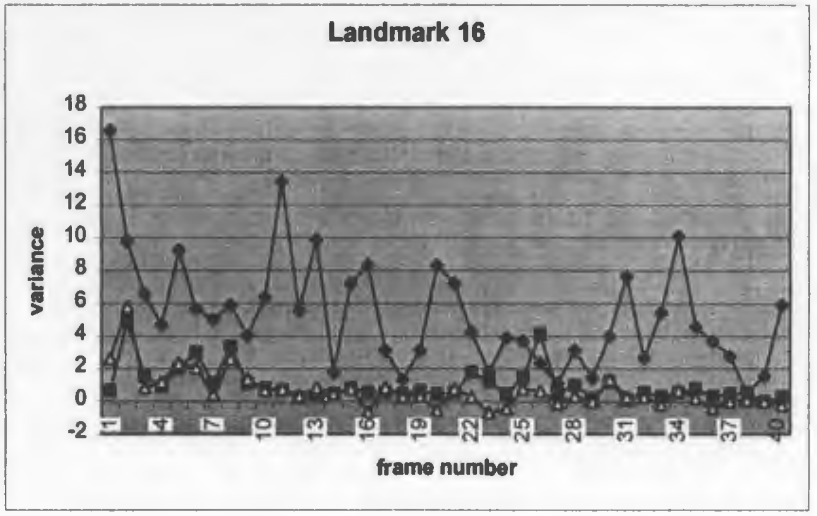


Appendix 5.2 Landmark Location Variation (for Landmarks 13 to 16)

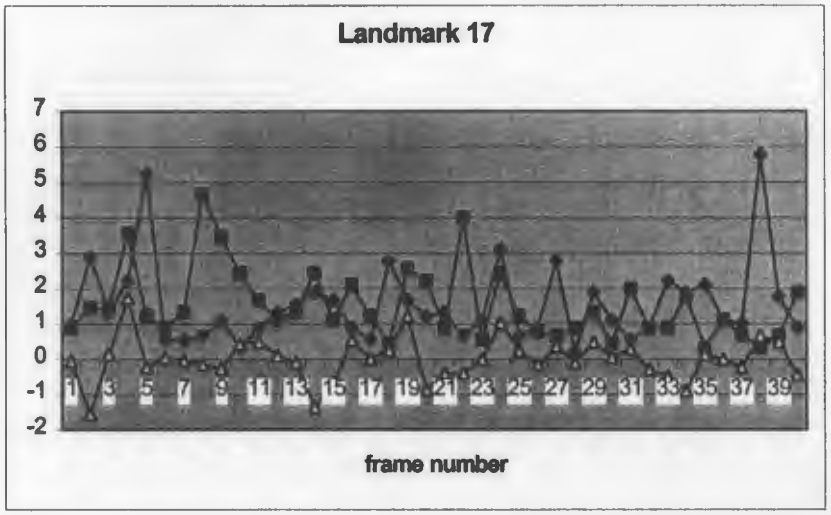


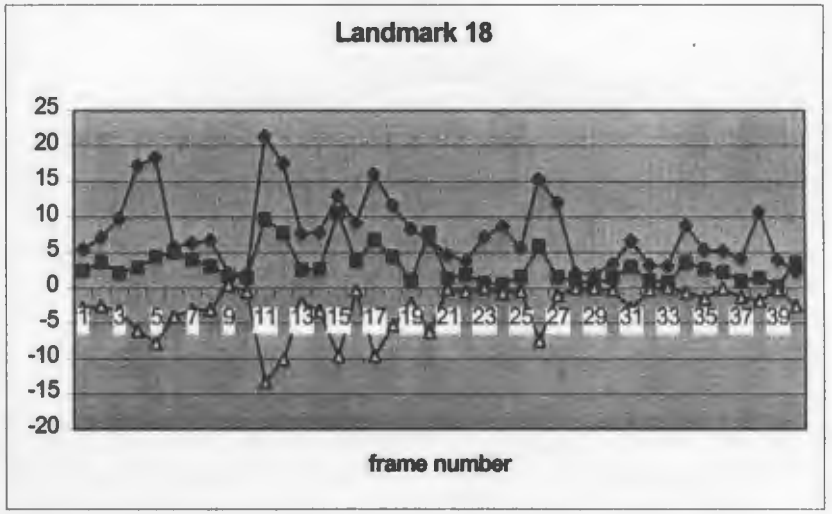


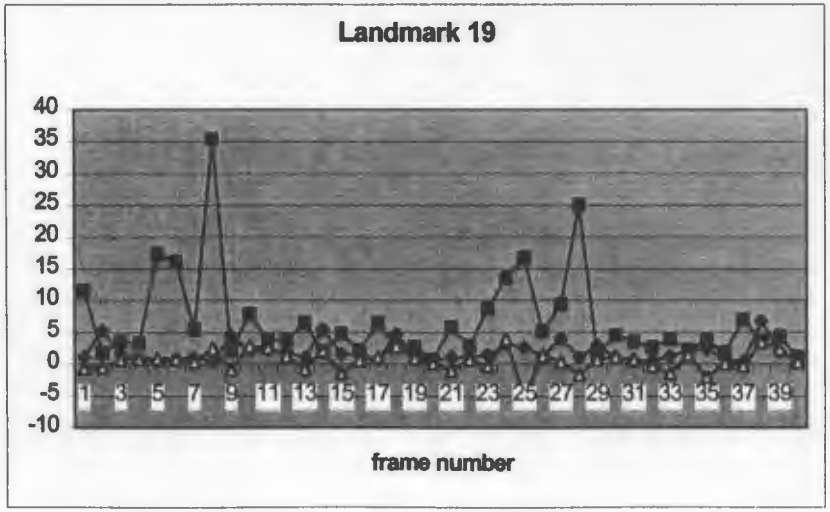


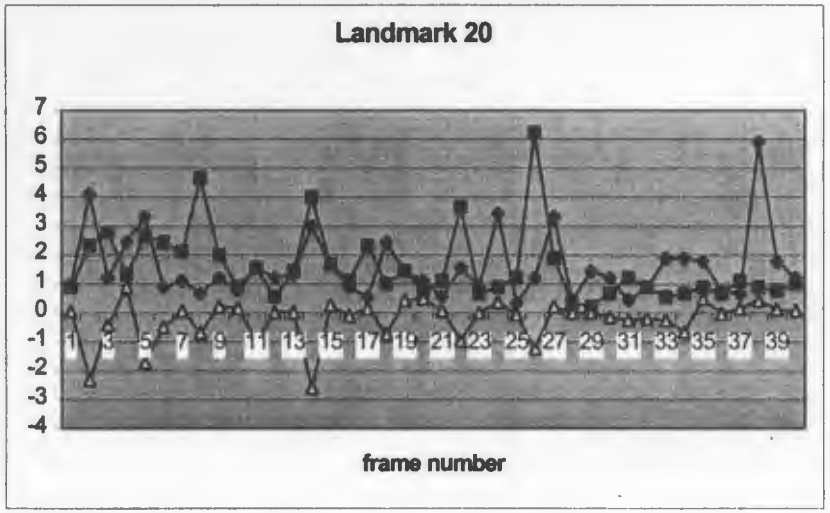


Appendix 5.2 Landmark Location Variation (for Landmarks 17 to 20)

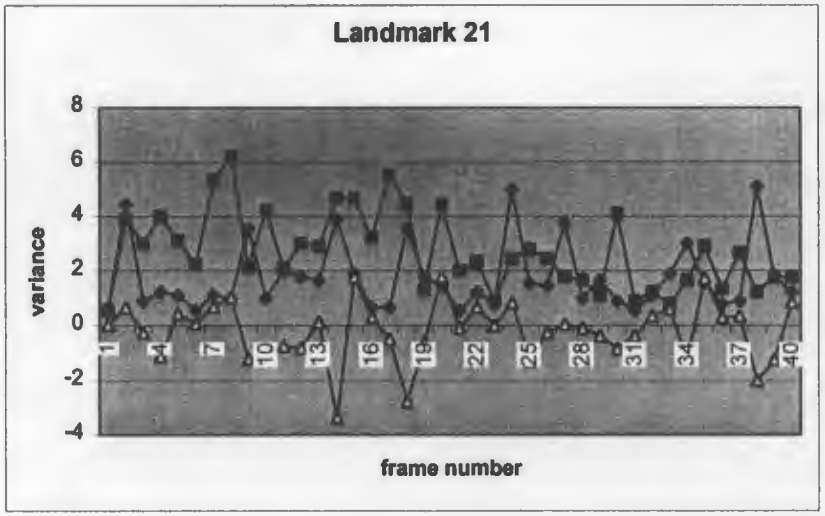


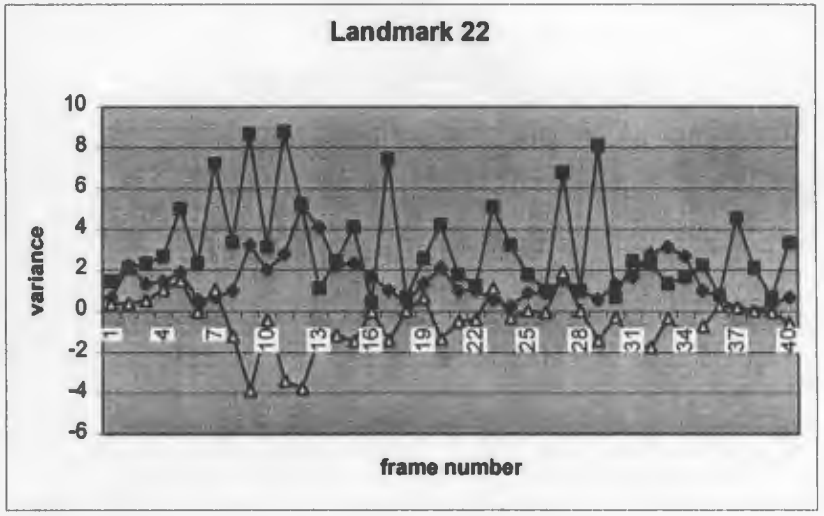


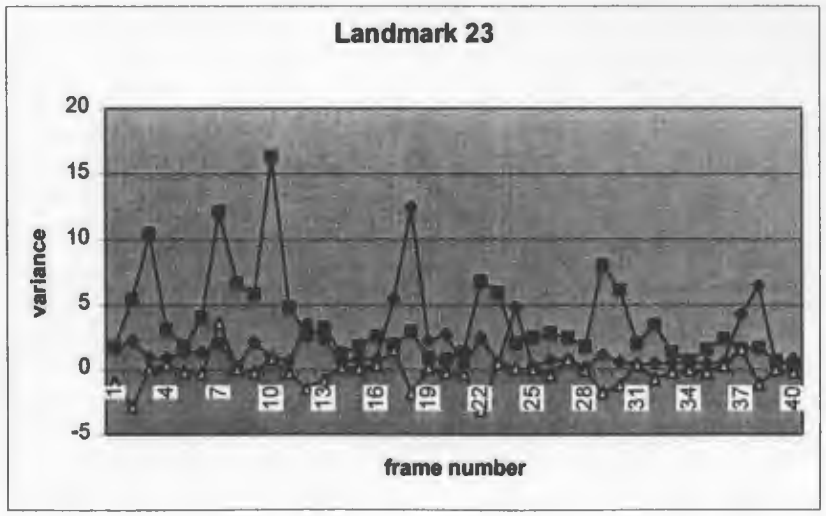


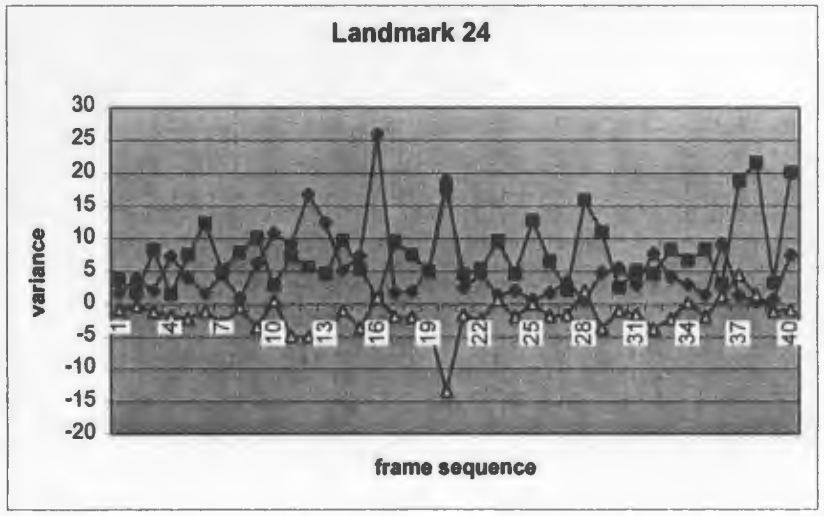


Appendix 5.2 Landmark Location Variation (for Landmarks 21 to 24)

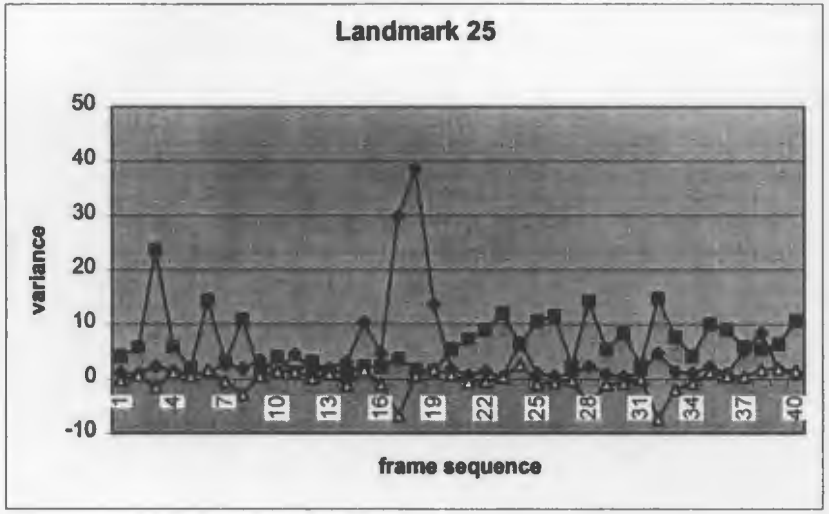


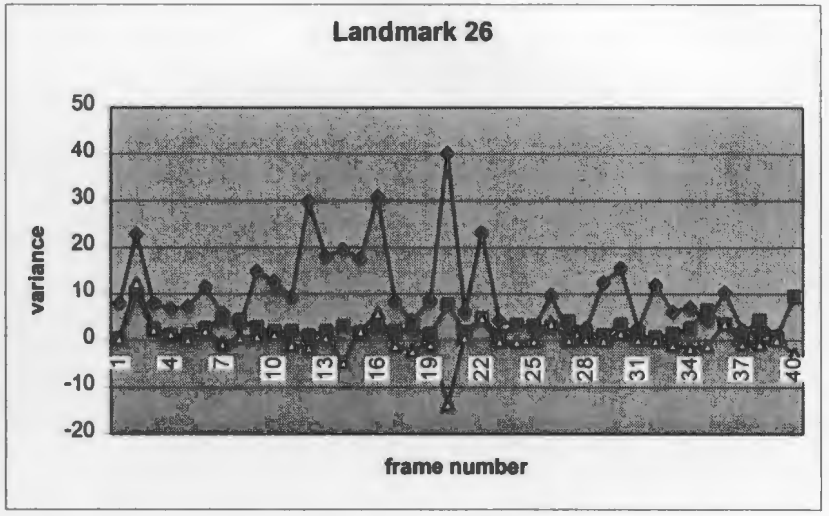


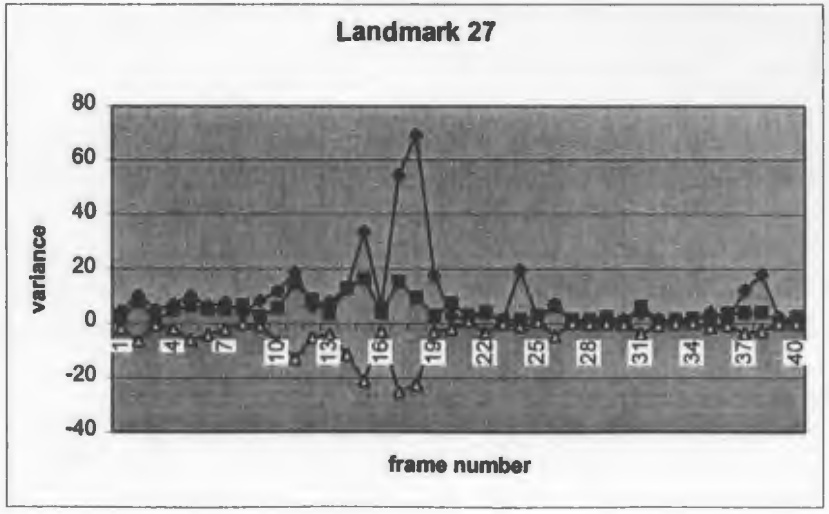


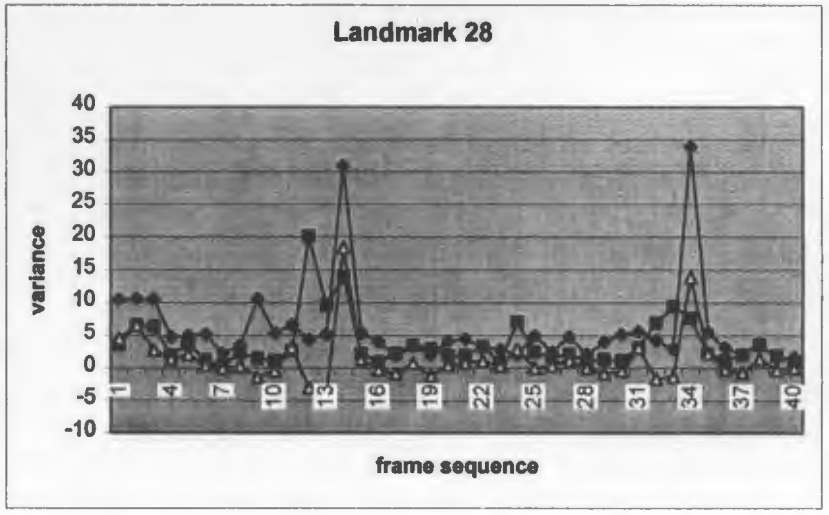


Appendix 5.2 Landmark Location Variation (for Landmarks 25 to 28)









Appendix 5.2 Landmark Location Variation (for Landmarks 29 to 30)